A PCA approach to the object constancy for faces using view-based models of the face.

by

Jevgenija Beridze

Department of Experimental Psychology

University College London

A thesis submitted for the Degree of Doctor of Philosophy

May 2020

Abstract

The analysis of object and face recognition by humans attracts a great deal of interest, mainly because of its many applications in various fields, including psychology, security, computer technology, medicine and computer graphics. The aim of this work is to investigate whether a PCA-based mapping approach can offer a new perspective on models of object constancy for faces in human vision. An existing system for facial motion capture and animation developed for performance-driven animation of avatars is adapted, improved and repurposed to study face representation in the context of viewpoint and lighting invariance. The main goal of the thesis is to develop and evaluate a new approach to viewpoint invariance that is view-based and allows mapping of facial variation between different views to construct a multi-view representation of the face. The thesis describes a computer implementation of a model that uses PCA to generate example- based models of the face. The work explores the joint encoding of expression and viewpoint using PCA and the mapping between view-specific PCA spaces. The simultaneous, synchronised video recording of 6 views of the face was used to construct multi-view representations, which helped to investigate how well multiple views could be recovered from a single view via the content addressable memory property of PCA. A similar approach was taken to lighting invariance. Finally, the possibility of constructing a multi-view representation from asynchronous view-based data was explored. The results of this thesis have implications for a continuing research problem in computer vision – the problem of recognising faces and objects from different perspectives and in different lighting. It also provides a new approach to understanding viewpoint invariance and lighting invariance in human observers.

Contents

Video files (in .avi format) representing non-rigid facial sequences defined in the thesis (included into dropbox folder)

List of Figures

1. Introduction

In the world, there is a widely held general belief that most people have incredible skill, which makes them experts at recognising faces. Some researchers claim that people are natural face recognition experts (Carey et al., 1992). Even though biometric authentication is now being used more and more to verify passengers at the airports, border guards still continue to match passengers' faces to their photo ID. Same as workers at the bank, or police, who also often verify people's identities using information received only from one photo. Evidence shows that the human visual system is highly effective in matching faces when it has prior information about those faces (Burton et al., 2011). Moreover, when it comes to familiar faces, humans can distinguish a face despite dramatic changes in illumination, viewing angle or partial occlusion (e.g. wearing sunglasses, scarf, hat, moustaches, etc.) (Hancock et al., 2000). This means that prior knowledge of the faces helps the human visual system effectively cope with problems arising from viewpoint and lighting changes in face matching tasks.

An understanding of what kind of processes within the human visual system help it to recognise faces under challenging conditions would support the development of effective recognition systems. For this reason, the current thesis aims to introduce a new PCA-based model approach of multiple appearances that will be used to explain object constancy for familiar faces in the human visual system. Two different theories have been used in this thesis to explain how the visual system in humans deals with facial discrimination from various viewpoints and different lighting. One theory of such discrimination is that we can bind different images of the same object into a single unit using a temporal association. (Miyashita, 1993) suggested that a close temporal association between newly viewed images in the sequence is sufficient to induce some *inferior temporal* (IT) neurons to respond similarly to arbitrary image pairs. Another theory introduced by (Marr & Nishihara, 1978), claims that the visual system is capable of storing and manipulating 3D object models. When a new 2D image is presented, the visual system determines whether it is possible to invoke the images of the 3D object that is already stored in the visual system.

This thesis favours a view-based approach over an object-centred approach as it has been inspired by humans' ability for facial recognition and has studied human vision. The motivation for the thesis was obtained from studies in psychology and cognitive science on the human visual system. Therefore, Chapter 1 will provide a brief description of the face processing pathway, with attention to the vital part of the brain responsible for face perception. It will also introduce the current state of the art that will explain the most important processes and abilities of the human visual system in terms of face constancy in various viewing conditions and define the essential features of a face used for face discrimination.

Chapter 2 will address the main strategies for encoding and animating faces in computer vision that are used today, together with the key methods used in this thesis.

As this thesis will apply an existing view-based approach developed by Cowe (2003) to build a computational model of the face that could cope with the viewpoint and lighting variations, this approach will be described in detail in chapter 3.

In chapters 4, 5 and 6, empirical studies will be presented that investigated the performance of the computational model and explored whether it is possible to reconstruct a new view of the face from another viewpoint in different viewing conditions.

Finally, Chapter 7 will provide some conclusions and suggestions for further research.

1.1. A psychological model of processing visual information in human vision

One of the most remarkable properties of human face recognition is the capability to discriminate faces in various scenarios. This capability is called object identification and is a process that works as a gateway from vision to cognitive processes in the brain (e.g. perception).

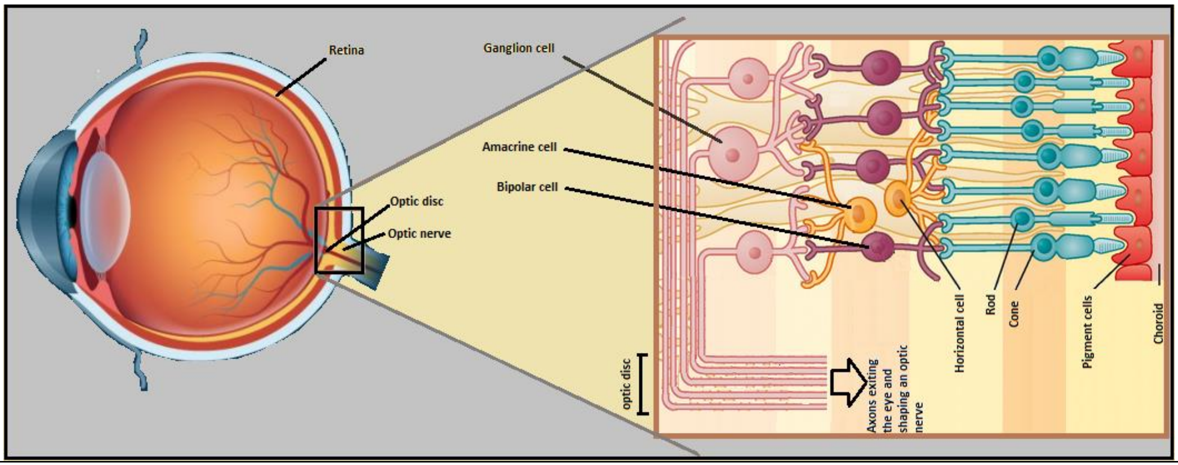


Figure 1 Diagram of the human eye, with a detailed layout of the retina and all its neuron layers in the human's eye. (image reproduced from Gary Heiting, OD. Retrieved from URL https://www.allaboutvision.com/resources/retina.htm; https://basicmedicalkey.com/the-special-senses-2/, Figure 8-2)

Visual perception in the brain starts as soon as the eye focuses light on the retina, where biochemical reactions are activated. The retina consists of five neuron types: photoreceptors, horizontal cells, bipolar cells, amacrine cells, and retinal ganglion cells. Photoreceptors consist of two types of cells called rods and cones. Due to rods, people can see in dimly lit places and observe motion, but only in black-and-white vision. Cone cells work in medium to bright light and are responsible for people's central vision and perception of the objects in colour. Photoreceptors are located at the back of the eye, so when the light reaches rods and cones, it is converted into an electric signal. This signal is then relayed via intermediate neurons (e.g. horizontal cells, bipolar cells and amacrine cells) towards the retinal ganglion cells (RGC). Then RGC, with their long parallel nerve fibres called axons, will help these signals to leave the eye. Axons create a bundle close to the optic disc (i.e. optic nerve head) and form an optic nerve that helps transmit visual information obtained from the eye to the brain (see **Figure 1**).

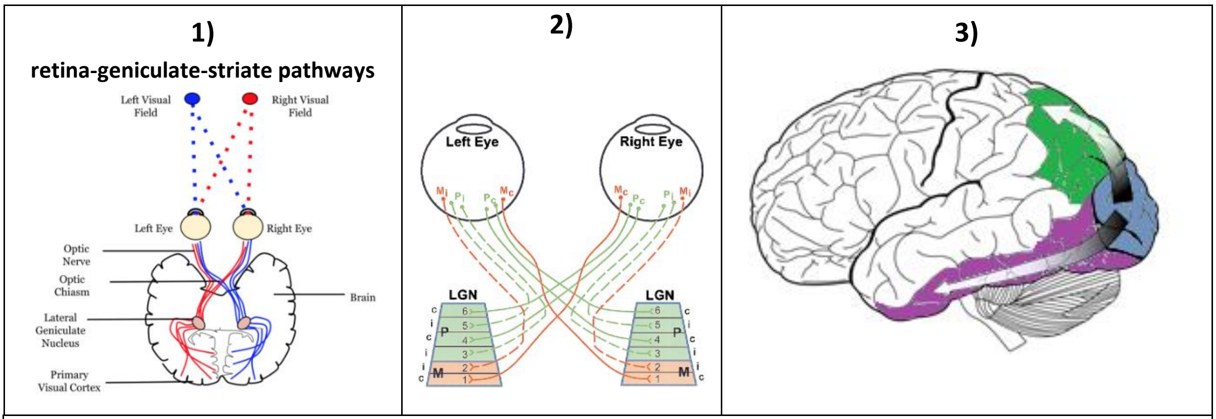


Figure 2 The transfer of visual information in the human brain. Image **1)** presents a model of transferring visual information to the primary visual cortex (V1) that is located in the back of the human brain (image reproduced from Wikipedia, user: Mads00). Image **2)** presents neuronal connections between human eyes and lateral geniculate nuclei (LGN) (image reproduced from URL https://www.researchgate.net/figure/Schematic-illustration-of-neuronal-connections-between-the-eyes-and-the-LGN-in-the_fig1_14059783 [accessed 6 Mar. 2021]). Image **3)** presents how V1 (grey) transmits information through two primary pathways of the visual system in the human brain. The dorsal stream is shown with green colour, and the ventral stream is presented in purple colour (image reproduced from Wikipedia, user: Selket).

From the retinal ganglion cells, the electrical signals travel along the retina geniculate striate pathways (see **Figure 2**, 1). Axons cross over in the middle part of the optic chiasm and pass through the lateral geniculate nuclei (**LGN**) found in the thalamus. They then continue all the way to the region of the brain called the primary visual cortex (V1), which is positioned in the calcarine sulcus that is located in the back of the human brain. Parallel nerve fibres (Axons) are divided into left and right sections that are found in both left and right hemispheres of the human brain. All inputs from the left visual field will be received by the right primary cortex (see **Figure 2**, 1), blue line), while the left primary cortex will receive inputs from the right visual field (see **Figure 2**, 1), red line). Within these pathways, there are a few more layers of neurons, this time located in the LGN (Meissirel et al., 1997): parvocellular layers, also called P layers (4 layers), and magnocellular layers (2 layers) or M layers. These layers of neurons are functionally distinct and are processed in different areas of the visual cortex. Neurons of P layers are anatomically smaller than neurons of M layers. Neurons of P layers are more responsive to colours, fine details or objects that are standing still or moving slowly, while M neurons respond more to object in motion (see **Figure 2**, 2).

In the primary visual cortex, a transfer of visual information is divided into two pathways: the dorsal stream and the ventral stream (Ungerleider & Mishkin, 1982). Ungerleider and Mishkin (1982) suggest that the dorsal stream is responsible for the perception of visual-spatial information and the ventral stream for visual object identification information (**Figure 2**, 3). Hence, the dorsal stream is involved in recognising where the object is in space, while the ventral stream helps to identify and form object representation. This is a brief description of the pathway of how visual information is transferred to the human brain. Next, we will define a vital part of the brain responsible for face perception.

1.1.1. Functional specialisation of the face-specific area in the ventral stream

A **Fusiform Face Area (FFA)** is a discrete brain region that is considered to be responsible for face perception in the human brain. It was initially described by (Sergent et al., 1992) and later deeper investigated and named by (Kanwisher et al., 1997). Kanwisher et al., in 1997, in their functional magnetic resonance imaging (fMRI) study, observed greater activation of this region in the middle fusiform gyrus (FG) when viewing faces as compared to objects. The FFA responded to human, animal or drawn faces much stronger than to the inanimate objects, body parts, back views of the heads, buildings, flowers, or scrambled images in healthy participants (Allison et al., 1999; Grill-Spector et al., 2004; Parvizi et al., 2012; Rotshtein et al., 2005; Tong et al., 2000) (see **Figure 3**).

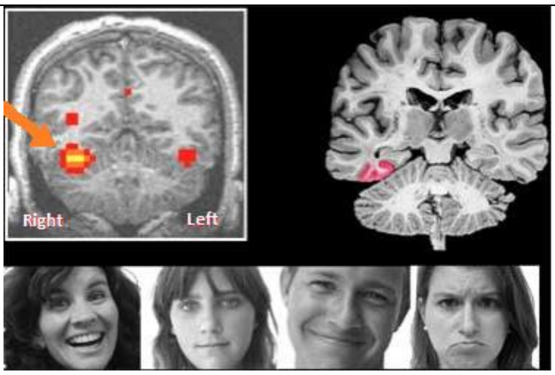


Figure 3 shows FFA activation presented in the fMRI scan (In the top left image, orange arrow). In the top right, there is an image of a post-mortem brain with FFA coloured in pink (image reproduced from autism-center.ucsd.edu).

An active discussion is ongoing if FFA may be a particular domain responsible specifically for face perception analysis, like face detection and recognition. Neuropsychology and studies on patients with lesions in this area suggest that such a specific place may exist. Due to brain specialisation, the loss of such a specialist area would selectively impair face processing and face identification. Dalrymple et al. in 2012, studied patients diagnosed with prosopagnosia, such people have problem with face recognition because they suffer from some form of pathological "face-blindness" when their natural face recognition skill is somehow missing or switched off (Behrmann & Avidan, 2005). Researches noted that these patients had a problem only with face identification while they were able to identify objects (Rezlescu et al., 2012). Thus they concluded that face processing might involve unique mechanisms that are different from those required for object processing. To have strong arguments in this discussion, researchers often use fMRI scanners that can easily present each region of the brain. Weiner and Grill-Spector, in their study, found two nerve clusters in the fusiform gyrus, entitled **pFus** and **mFus,** which had a stronger response to faces than to objects or body parts (Weiner & Grill-Spector, 2010). This result suggested that the FFA could be composed of functional clusters that were at a finer spatial scale than what has been measured previously. Parvizi et al. continued this work and showed that mild electrical intracranial stimulation of these functional clusters (pFus and mFus) could cause visible distortions in face perception (Parvizi et al., 2012). In 2014, Rangarajan et al. brought evidence of functional brain lateralisation for face processing (Rangarajan et al., 2014). In their intracranial study, electrical stimulations were applied to the right hemisphere face areas, which caused distortions in face-selection. When similar stimulations were used on the left hemisphere, they did not affect face perception but caused non-face, simple distortions like phosphenes. These results suggest that there is some innate face-selective area in the human brain that is dedicated only to face processing.

On the other hand, Gauthier and Logothetis (2000) tried to prove that the FFA was an expertise area rather than a face-processing area. Their study tested both bird and car adepts and found small activation in the FFA area when bird adepts were recognising birds and car adepts were recognising cars. These results were replicated by Xu and McGugin et al., confirming that the effect of expertise was found in the FFA area (McGugin et al., 2012; Xu, 2005). Similar results were further demonstrated in other studies that used chess displays (Bilalic et al., 2011) or studies with faces presented in x-rays (Bilalic et al., 2016). Nevertheless, the discussion about the nature of the functionality of the FFA role remains open.

The next subchapter will present the current state of the art in the research on the processes and abilities of human vision in terms of face constancy for familiar faces.

1.2. Object constancy for faces

Face perception is a complicated process performed by the human visual system that is essential to coexist in society. Many types of signals and information are visible on a human face, but processes and methods that extract this information are complicated. Faces differ from each other by shapes and sizes; moreover, they are covered with a large number of muscles, which adds diversity and complexity to the face (Bruce & Young, 1998). These unique variations of the face distinguish one face from another, and humans are capable of determining individuals by their faces effortlessly (Hancock et al., 2000).

However, face recognition can be challenging because the appearance of a face can alter a lot in different viewing conditions (DiCarlo et al., 2012; Johnston et al., 1992; Lander et al., 1999; Logothetis & Sheinberg, 1996; Pinto et al., 2008; Poggio & Ullman, 2013; Tanaka, 1996). The various conditions may include object direction, illuminance or variability (size, colour, and other category differences). One of the main difficulties that stop people from recognising two views of individual of being the same person is when this individual is unfamiliar to them, and they had no previous experience of seeing those two views at the same time (Natu & O’Toole, 2011; O’Toole, 2005, p. 360).

The thesis will address these limitations of the human vision in chapter 6, where the computational model will be created to investigate face constancy for unfamiliar faces.

Although recognition of unfamiliar faces can be challenging, people can effortlessly recognise an object of interest under various conditions if they had prior knowledge of that object. This means that people have the skill to achieve object constancy. In psychology, object constancy is defined as a perceptual phenomenon when characteristics of an object appear to be perceived the same regardless of changes in viewpoint, presentation, lighting or distance from the observer. As we know, humans have this perceptual phenomenon, and the human visual system can easily disregard the problems that appear due to changes in pose or lighting. Does this mean that the human visual system achieves object constancy because it is view or lighting invariant?

The scientific evidence suggests that this is not the case. An inversion effect is one piece of evidence that supports view dependency in the human visual system. Faces are harder to recognise when they are upside down than when they are upright (Farah et al., 1998, p. 482; Yin, 1969, 1970). This evidence was used to argue that the FFA has an orientation dependency as a result of human experience being overwhelmingly biased towards upright faces. Thompson, in 1980, demonstrated this view dependency by creating a so-called Thatcher illusion **Figure 4**, where the eyes and mouth of an expressive face were excised and inverted (Thompson, 1980). When the face is upright, the resulting image looks grotesque. However, when the face is inverted, it is hard to note that there is anything different about the face. The Thatcher illusion is a perfect example of how face alignment affects human perception of faces. Also, the results of this experiment suggest that the human vision processes face holistically, rather than based on the single features on the face.

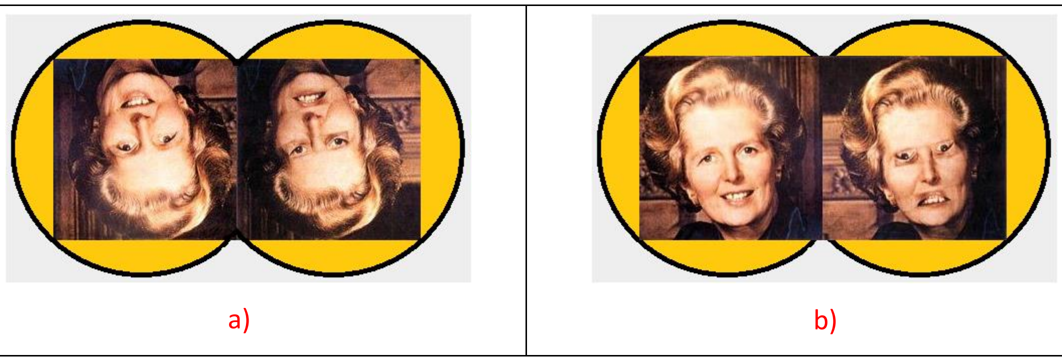


Figure 4 The Thatcher illusion. (a) Ms Thatcher's head positioned upside-down; it is difficult to notice that the eyes are placed in the reverse direction in the right image. In (b), the face rotated upright; here, we can notice a strange face appearance (image adapted from Thompson 1980).

Also, several psychophysical studies presented below confirmed view dependency of the human visual system.

Hill et al., 1997, in their study, confirmed that participants performed poorly in face matching task when viewing positions were changing. However, once participants learned all faces, recognition performance increased, and all the faces were matched well. In this study, participants turned to be quite poor at generalising a face across different views from a single face view. The decrease in performance was directly related to an increasing difference between the views. (Van der Linde & Watson, 2010) also noted that the highest face recognition performance is achieved when an angular offset between learned and tested face image is the least, while massive changes in yaw rotation are the most damaging. However, if humans would systematically experience changes in viewpoint directions, it would have a high impact on the process of face recognition.

Watson et al., on the other hand, found a difference in viewpoint-dependency for rigid and non-rigid face motions. Their study showed that non-rigid facial transformations appeared to be less viewpoint-dependent than rigid motions of a face. According to Watson et al., "it seems that the currently prevailing theories of object recognition will in the future need to account for not only patterns of static object recognition but also those of object motion" (Watson et al., 2005).

The human vision is also lighting dependent. Hill and Bruce, in their study of comparison tasks, used data of faces scanned with a laser and presented in various lighting and viewpoint conditions. They found that variations of light on a face (especially bottom lighting) can cause difficulties for the face matching performance (Hill & Bruce, 1996). Similar results were obtained by (Braje et al., 1998) (see **Figure 5**).

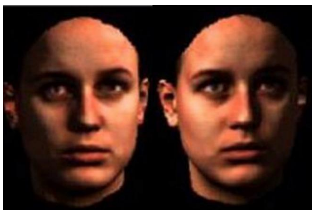


Figure 5 Lighting stream applied the left and right side of the face. These two images demonstrate the kind of lighting used in an experiment by Braje et al. (1998). After showing the image demonstrated on the left, participants were very accurate in determining whether another image such as the one on the right presents the same or a different individual (in this case the same) (image reproduced from Braje et al. (1998)).

Also, Braje in 2003 discovered that illumination directed from the above gives advantages in the comparison tasks (Braje, 2003). This finding is coherent with evidence that says that the observer needs only one light source positioned approximately over the head to recover simple shape from shading patterns (Ramachandran, 1988). In addition, Johnston et al. (1992) found that if the light was directed from below and then the face was inverted, it was possible to reduce the effect of the face inversion for full-face views (see **Figure 6**).

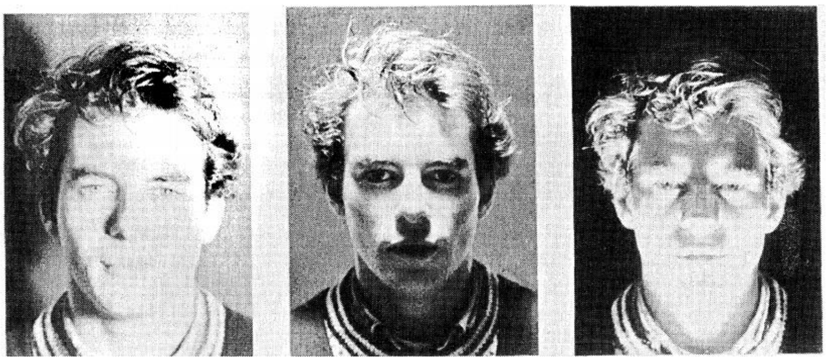


Figure 6 presents a face in different lighting condition. These images were used by Johnston and colleagues Johnston et al., 1992) in an experiment to define if these face images belonged to the same participant. The image on the right represents a "ghost-like" face appearance which is harder to recognise than the rest of the images in the same row. (image reproduced from Johnston et al. (1992))

All these findings on luminance directions provide strong evidence of some face encoding scheme. However, hollow face illusion, when a concave face mask appears to be convex, showed that the direction of the light source was not the only parameter that disrupted recovering of shape-from-shading (Hill & Bruce, 1993). According to Hill and Bruce, familiarity with the 3-D depth structure of a face was "an important component" in the process of perceiving an illusion.

What about the size, does it impair face reconstruction in the same way as changes in viewpoint or lighting? Lee et al. (2006) used unfamiliar faces that differed in size up to fourfold and found that face matching performance is size invariant. Recently, Guo et al. (2013) investigated how a pose of face interacts with size changes. In their study, participants were very accurate in matching faces, regardless of the face sizes ranging from arm's length to up to five metres. However, when the viewpoint was changed, the participants' ability to match faces decreased drastically. Evidently, the size does not significantly impact the face matching tasks as opposed to presenting a face from different viewpoints.

We know that a face in one pose will look different from different angles and project many different patterns onto the retina that may challenge human vision in obtaining viewpoint independence and finally achieving object constancy. Then, how is human vision capable of achieving viewpoint independence for faces while being so view-dependent?

Physiological experiments on the macaque monkey brain demonstrate that object recognition in the cortex is performed via the ventral stream, which carries visual information from the primary visual cortex (V1) to the secondary visual cortex (V2), continues to the visual cortical area (V4) and finally reaches the inferotemporal cortex called IT (Ungerleider & Haxby, 1994). The inferotemporal cortex is believed to be a fundamental component in object recognition. In IT, researchers found cells that are tuned mainly to the views of complex stimulus like faces. In an experiment, these cells had a strong activation response to face pose and little or no activation to objects such as bars, spots, edges, hands or brushes (Bruce et al., 1981). Neuropsychological studies, together with fMRI studies, agree that the inferotemporal cortex is an essential component for object recognition also in human vision (Logothetis & Sheinberg, 1996; Ungerleider & Haxby, 1994). While moving up the visual hierarchy, the neurons increase receptive field sizes with the increasing complexity of stimuli. Neurons in V1 have small receptive fields and respond only to simple stimuli such as moving lines or bars (Hubel & Wiesel, 1962). They are also selective to orientation and very dependent on the contrast of the stimulus (Polat et al., 1998). Whereas IT neurons have larger receptive fields than V1 neurons and, as previously mentioned, will respond selectively to complex stimuli such as objects and faces (Bruce et al., 1981; Desimone et al., 1984; Tsao & Livingstone, 2008). In other words, receptive fields are becoming more invariant to position and specific to structure as visual information proceed up the ventral stream from V1 to IT (Kobatake & Tanaka, 1994).

In addition, researchers Peret and Oram, in their study, presented that viewpoint invariance to any transformation can be achieved with a hierarchical model where outputs of cells that respond to different views of the same object are being pooled over to build up one object that is seen from different views (Perrett & Oram, 1993). Poggio and Edelman, a couple of years earlier, also proved that achieving viewpoint-invariance can be established with pooling operation (Poggio & Edelman, 1990). They trained a learning network called Gaussian Radial Basis Functions (GRBFs) that used a small set of object views (paperclip-like objects) to achieve object constancy for that specific object at any angle. Their approach showed that novel invariant views of the object are achieved by interpolation between already stored ("learned") views in the learning network. Psychophysical studies (Bülthoff & Edelman, 1992; Logothetis et al., 1995; Tarr, 1995) together with physiological studies (Booth & Rolls, 1998; Kobatake et al., 1998; Logothetis et al., 1995) discovered that the learning process forms IT neurons, which are tuned to full or partial views and found several neurons that are selective for view-independent representations (Perrett et al., 1991).

Additionally, several researches found neurons that responded to a wide range of rigid head motion in yaw rotation (Desimone et al., 1984; Hasselmo et al., 1989). Perret et al. defined five neurons, each of them was responding to different but only one position of the head (e.g. to full face, profile, rare view head, head up or head down) (Perrett et al., 1992). These pieces of evidence seem to suggest that the brain is storing face frames in the two-dimensional face space. This way, every view of a face is collected and encoded separately in order to perform face discrimination in multiple viewpoints and in various viewing conditions. This evidence became one of the key reasons why this thesis chose to focus on a view-based approach to build a PCA-based model of multiple appearances. The multiple appearances model can use various views of a face and refer them to a single general cause (e.g. smile) using a multi-view vector. This multi-view vector will then be able to present how the smile will look from a different perspective. In this way, one can get a representation by underlining the cause of expression in its various appearances. It is important to mention that the multiple appearances model is not a single appearance model, despite being composed of one multi-vector. Within this model, different appearances will be grouped to represent the same thing, making the multi-view vector an excellent mathematical alternative that reflects the biological processes in the human brain, where every view of a face is collected and encoded separately but represents the same face. A detailed description of this model will be given in chapter 4.

1.3. Chapter 1 summary

Face perception is a remarkable human ability and one of the most complicated tasks performed by the human visual system. People, having prior knowledge of a face, can easily distinguish a face despite dramatic changes in illumination or viewing angle. Hence, they can cope with those type of viewing variations in face matching tasks and achieve invariance in viewpoint and lighting. The information processing pathway for matching faces starts at the retina and runs through optic chiasm over the optic nerve, passing the lateral geniculate nuclei (LGN) until it reaches V1. Then it travels into the inferotemporal cortex (IT) containing the FFA face-specific region, where the face is finally perceived. The inferotemporal cortex (IT) is a fundamental component in object recognition. In IT, researchers found cells that are tuned mainly to the views of complex stimulus like faces. Moreover, psychophysical and physiological studies found that IT neurons are being formed through a learning process and are tuned to full or partial views. Besides, they found only several neurons that were selective for view-independent representations in the inferotemporal cortex. In addition, the human visual system process faces holistically rather than based on the single features on the face.

Furthermore, some neuropsychological studies suggest that the human brain is storing human faces in a 2D format. The brain collects and encodes every view of a face separately. For this reason, the current thesis has chosen to use a 2D view-based approach to build a multiple appearances computational model to explain object constancy for faces. The topic of face constancy in various conditions, received significant interest in recent years, especially in the field of computer vision that aims to create realistically-looking computer animations of a face. As stated previously, this thesis is oriented to build a computational model of a face. Hence it will require to incorporate computer facial animations to create lifelike representations of a face. Therefore, the next chapter will discuss approaches on how to build the realistic reconstructions of a face that will be used to construct the computational model.

2. A computational model of a human face

As previously discussed, most people have incredible skill, which makes them experts at recognising faces. Humans can define if a face is familiar or unfamiliar, even under large rotations in the viewing angle or the very complicated lighting conditions, by merely looking at someone's face. Psychological evidence showed that people transfer information obtained from the carrier's face to the brain, where it performs cognitive processes. Thus, the human's brain may be defined as a computer that runs vision algorithms to obtain object constancy for faces.

Hence, this chapter will present the main strategies for encoding and animating faces in computer graphics addressing their weaknesses and relation to psychological theory on encoding the dynamic change in faces in the brain. Also, the chapter will introduce key methods used in this thesis that will address these weaknesses based on natural processes that are happening in the human visual system described in Chapter 1.

2.1. Computer facial animation

Computer facial animation is a part of computer graphics that aim to create realistic animated representations of the head and face. It uses mathematical modules together with computer vision techniques to mimic realistic movements of both internal and external features of a face (e.g. facial expressions, head shape, hair and so on). The first question on how to generate realistic animated faces was raised several decades ago (Parke, 1972). However, this task is still very challenging, and there is still no clear answer. The challenges are due to complex mathematical calculations that are used to encode the sophisticated structure and multi-dimensionality of a face and the computational speed of hardware that is required to run all these calculations. This thesis is interested in generating a realistic representation of a face, which is based on natural processes that take place in the human brain. From the psychological theory introduced in chapter 1, we know that even though the human face is built in a three-dimensional shape, the brain processes the human face in a two-dimensional shape. Therefore, this subchapter will cover some methods of three-dimensional face modelling that will incorporate two-dimensional representations of a face.

2.1.1. A 3D face modelling approach

2.1.1.1. Polygons

Frederic Ira Parke invented the first 3D face model in 1972. He digitised it by hand by merely drawing polygons (triangles connected at each vertex) on a real human face and then collecting several frames of pairs of orthogonal views of the face in different expressions. Next, he manually measured vertexes (key poses) from photographs and used a linear interpolation on the three-dimensional position of the vertexes to obtain a non-rigid motion. Unfortunately, his model had several shortcomings. First, the model required some artistic skills to draw polygons on the face manually. Second, since faces have different shapes and sizes, the key poses taken from one face will not have a perfect fit on other face shapes. And finally, the whole process was very time-consuming.

Already in 1987, Rydfalk introduced an automatic approach called Candide that created the most realistic 3D face model without computational complexity. This 3D face model used a small number of polygons to encode the human face (75 vertices and 100 triangles), which allowed a fast face reconstruction with relatively low hardware ability. Even though the created approach was fully automated with a short processing time, the generated face model had very edgy facial features that made it look unreal. This approach was not able to realistically reconstruct the volume of face shape and facial features (e.g. eyes, nose, cheeks, lips); hence face model lacked the complexity present in the real human face. (see **Figure 7**).

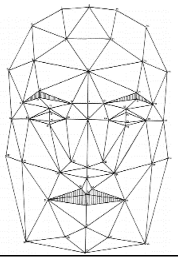


Figure 7 Candide 3D face model consisting of 75 vertices and 100 triangles (image reproduced from (Rydfalk, 1987)).

In the same year, Waters (1987) introduced a parametrised muscle approach to realistically reconstruct complex face structure in facial animations (see **Figure 8**, a). His approach also used polygon meshes together with *The Facial Action Coding System* (FACS) model. This model was created by Ekman and Friesen (Ekman & Friesen, 1978) and could define emotional states from the visible changes in the face called Action Units (AU). AU can consist of one or a group of muscles to obtain realistic movement of the skin in the face. Waters implanted ten AU in his face model. These AU were responsible for driving specific zones in the face, their key nodes in the facial model displaced with a circular cosine falloff function. Waters model became a foundation for physically based rendering in facial reconstruction. Later, Terzopoulos and Waters upgraded Water's model by adding anatomical structure and facial dynamics, by adding three layers of deformable mesh that were responsible for the movement of skin elasticity (cutaneous tissue), fatty tissue (subcutaneous tissue) and muscle layer (Terzopoulos et al., 1993). Elastic springs connected each key node of deformable mesh at each layer. These elastic springs were adjusted with different non-linear stiffness values, separately for each layer, to obtain different properties of a face from these sub-layers. This model gave very realistic results in deformations of skin in facial reconstructions and is broadly used in the film industry. For example, ***Weta digital*** used this model to create and move all Na' vi characters in Avatar's movie, directed by

James Cameron (Wikipedia contributors, 2019) (see **Figure 8**, b). Even though this model gave significant results in realistic face reconstruction and related to psychological theory, the biggest drawback of this method is that it required very complex computations and high-performance hardware to render these facial animations.

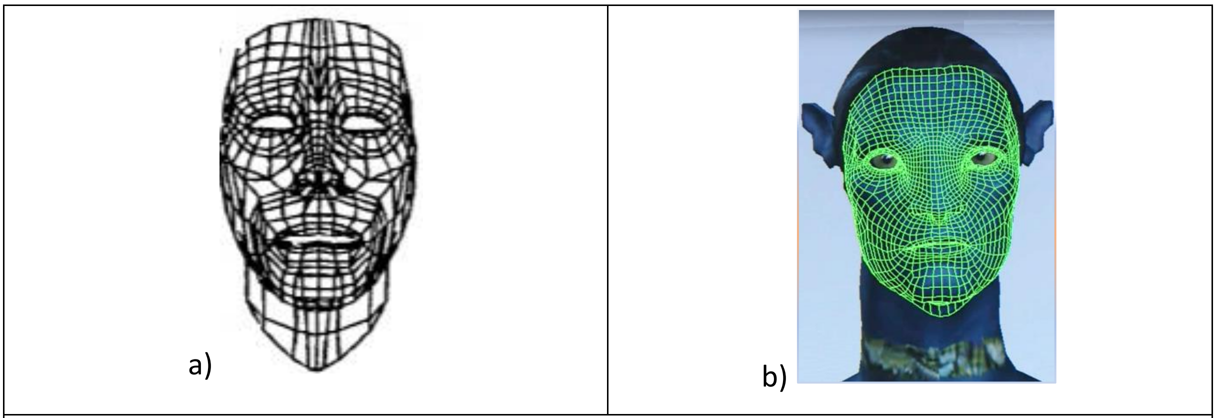


Figure 8 a) Figure shows Waters polygonal mesh method (image reproduced from (Waters, 1987)). **b)** The figure shows how Na'vi characters from the Avatar movie were composed of polygonal meshes based on three layers deformable mesh model (image reproduced from (Cameron, 2009)).

2.1.2. Image-based approach

Another technique that also relates to psychological theory, but requires slightly fewer computations and resources to build a realistic facial animation than the previously mentioned approaches, is the image-based approach. This approach uses only two-dimensional image information as a keyframe to construct a three dimensional morphable model of a face.

DeCarlo & Metaxas (1996) significantly contributed to the development of image-based techniques. They used the morph technique to blend the texture from two-dimensional images onto a three-dimensional face model. They were interpolating between keyframes to find a proper three-dimensional position to achieve realistic facial animation. On the other hand, Pighin et al. (1998), in their approach, used a combination of both face shape and texture information. They created textured face meshes for every keyframe. These keyframes were linearly interpolated to achieve facial animation, then blended and wrapped onto a three-dimensional model. For both approaches, the image-based method gave realistic results. Still, both of them encountered a registration error that is the main disadvantage of these approaches because it affected the texture of the face making it slightly blurred.

A method that handled this registration error was first introduced by Blanz and Vetter (1999)**.** Their method used a dataset of 200 3D facial laser scans in various shapes and non-rigid movements to build a photo-realistic 3D morphable face model. Each face image contained approximately 70.000 3D vertices and an RGB texture map. Image scans had to be pre-processed to remove global 3D transformations. Image registration of the vertices was based on the optic flow method. 3D morphable face model automatically learned class-specific face information from the collected dataset, such as changes in viewpoint and illumination. The model used face shape and texture information extracted from images as vectors but analysed them separately in a statistical technique called Principal Component Analysis. The PCA technique successfully reduced the dimensionality of the model's dataset. Model’s shape vector is defined by (2.1)

$$\boldsymbol{S} = (\boldsymbol{x_1}, \boldsymbol{y_1}, \boldsymbol{z_1}, \dots, \boldsymbol{x_n}, \boldsymbol{y_n}, \boldsymbol{z_n})^{\boldsymbol{T}} \qquad (2.1)$$

Where parameters **x, y, z** present pixel coordinates. Texture vector is defined by (2.2)

$$\boldsymbol{T} = (\boldsymbol{R_1}, \boldsymbol{G_1}, \boldsymbol{B_1}, \dots, \boldsymbol{R_n}, \boldsymbol{G_n}, \boldsymbol{B_n})^{\boldsymbol{T}} \qquad (2.2)$$

Where **R, G, B** present colour values. First, they computed an average value for shape and texture ($\overline{\boldsymbol{S}}, \overline{\boldsymbol{T}}$). Then found eigenvectors ($\boldsymbol{S_i}, \boldsymbol{T_i}$) and eigenvalues ($\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}$) that were used to define coordinates of face images in the PCA face system. PCA face system is defined by (2.3)

$$\boldsymbol{S_{mod}} = \overline{\boldsymbol{S}} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n-1}} \boldsymbol{\alpha_i} * \boldsymbol{S_i} \; and \;\; \boldsymbol{T_{mod}} = \overline{\boldsymbol{T}} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n-1}} \boldsymbol{\beta_i} * \boldsymbol{T_i} \qquad (2.3)$$

Where shape and texture models $\boldsymbol{S_{mod}}$ and $\boldsymbol{T_{mod}}$, are formed by $\boldsymbol{S_i}$ and $\boldsymbol{T_i}$ eigenvectors, and $\boldsymbol{\alpha_i}$, $\boldsymbol{\beta_i}$ eigenvalues. (see Formula 2.(2.2)). **Figure 9** shows an average face and variations of shape and texture based on principal components (eigenvectors) with added eigenvalues.

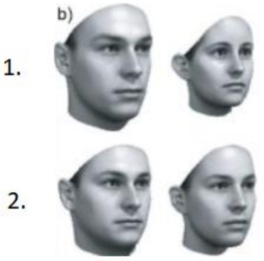
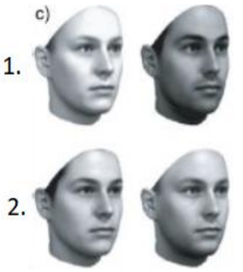
Average face	**Shape**	**Texture**
	 $-3\sigma \quad +3\sigma$	 $-3\sigma \quad +3\sigma$

Figure 9 shows an average face (a). Figure b) and c) shows a first two principal components (eigenvectors) of a data set in PCA face space with -3 and +3 standard deviation from the average face. b) The figure shows how face shape is changing with different eigenvalues while texture stays constant. c) The figure shows how the texture of a face is changing while keeping a face shape constant (image adapted from Blanz & Vetter (2003)).

The previously mentioned Blanz and Vetter approach could generate a realistic representation of the 3D face model that achieved pose and lighting invariance using only one image example of the face. However, to build a model, they used an image that contained a whole representation of a human face. Since humans cannot see all views at the same time, we will now describe image-based methods that investigated face reconstruction under changes in the pose using face information taken only from one perspective.

First 2D image-based method that synthesised a face in different perspectives was based on piecewise warping. The piecewise warping approach transforms the shape of the face image in a piecewise manner to another specified pose. Beier et al. presented a feature-based technique called morphing that smoothly transitioned one image into another by using an interpolation process between two images (Beier & Neely, 1992a). First, the morphing process warps two face images so that they have the same "shape", next it cross-dissolves images to blend one image into another smoothly. In the warping process, images are aligned with lines that relate to positions established in the source and destination images. These lines are moved with high precision to where they are mapped to, and everything else is blended smoothly based on those positions. The biggest drawback of this method is that these positions are settled manually. In addition, this method had a problem with speed and control because it is a global method and all line segments needed to be referenced for every pixel. Sometimes, an algorithm was generating unexpected interpolations as it tried to guess what should happen far away from the line segments. This problem usually demonstrates itself as a "ghost" of a part of the image showing up in some unrelated part of the interpolated image. In the empirical study of chapter 4, we will compare the face reconstruction performance of this morph algorithm with the reconstruction performance of our method.

One of the first image-based multi-view methods that handled pose variation in face recognition was suggested by Beymer and Poggio (Beymer & Poggio, 1995). In their approach, they used prior knowledge of a face to synthesise a face in new perspectives because prior knowledge integrated into the system improves face generalisation across views (Niyogi et al., 1998). Prior knowledge was one available real single view of a face that was used to synthesise this face in other poses. Beymer and Poggio introduced a parallel deformation method to synthesise a set of rotated virtual views from one real example view of the person. After estimating a pose of the face, the algorithm automatically found the first three points and then shifted the test images to match the candidate poses from the database. This method used a gradient-based optical flow algorithm to find displacement fields that collected changes on a pixel level between two poses of a face. These displacement fields were registered as a template displacement field (see **Figure 10**, (prototype flow)).

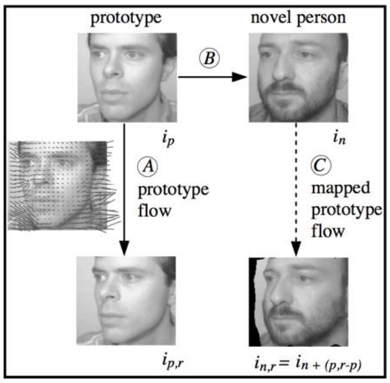


Figure 10 Method of parallel deformation. (A) with optical flow prototype flow shows changes between images $\boldsymbol{i_p}$and $\boldsymbol{i_{p,r}}$, (B) next, the prototype flow is applied onto the novel face $\boldsymbol{i_n}$, and (C) the novel face image is mapped to the virtual view (image reproduced from Beymer and Poggio 1995).

Next, linear combinations estimated a template displacement field of a given testing face image. After learning a displacement field, it was used to generate a new image under different poses.

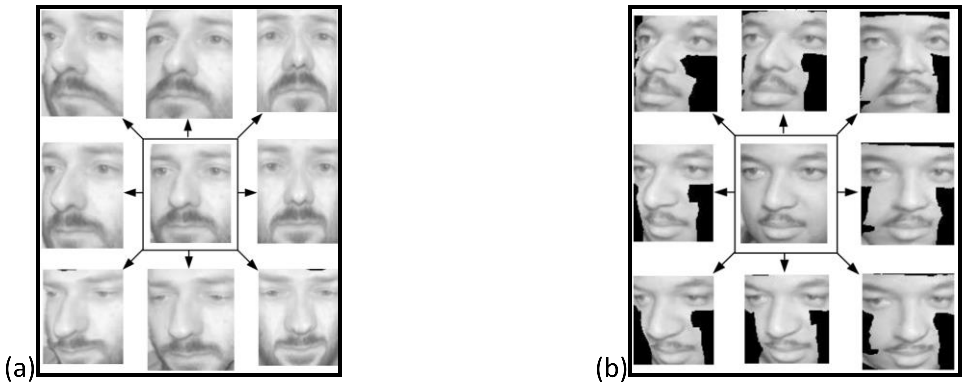


Figure 11 Synthesis of virtual views on the generated shape. In (a), face mapping over multiple poses was made manually. In (b), face mapping was accomplished automatically with an image vectorizer (image reproduced from Beymer and Poggio 1995).

Their method performed manual technique and auto technique to generate new virtual views. Both techniques generated eight virtual views per person from one example view in about 15-degree rotation away from the original pose (see **Figure 11**). These face poses were ranging from -30° to + 30° (yaw) and from -20° to +20 ° (pitch). The database consisted of 62 participants, where each participant was presented in 10 different poses. The method obtained better face reconstruction results with a manual technique (82 per cent) than with an auto technique (75 per cent).

Although this method meets the requirements of achieving object constancy represented in psychology theory and reproduce an authentic reconstruction of the complex structure of a face, it contains several significant shortcomings. First, this method performs very complicated calculations to find a relation between images. Second, this approach requires high data storage per participant. Third, it fails to reconstruct a face if the yaw angle between two poses is too large or some regions of the face become invisible due to self-occlusion, which leads to the loss of information on the face (see **Figure 11**, (b)). Moreover, it is unclear if this approach could be reused to work with nonuniform illumination conditions.

As we may observe, face reconstruction based on 3D methods appeared to be one of the most successful strategies for handling large pose variation and various lighting conditions. Thus, Blanz and Vetter reused their previous 3D approach (Blanz et al., 1999) to synthesise a new face from a single 2D face image (Blanz et al., 2002; Blanz et al., 2005; Blanz & Vetter, 2003). They described a synthesised face by a vector that used a combination of weighted functions with carefully chosen coefficients for both shape and texture in order to obtain as similar as a possible reconstruction of a face with an input image (see **Figure 12**).

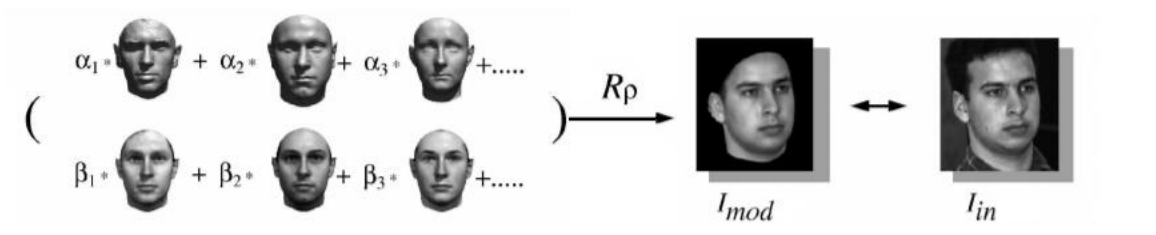


Figure 12 shows how the fitting process is accomplished in Blanz and Vetter 3D face model. In their fitting process, they carefully choose coefficients for both shape ***α*** and texture ***β*** to obtain a representation of a face $\boldsymbol{I_{mod}}$ that would be similar as possible to the input image $\boldsymbol{I_{in}}$. (image reproduced from Blanz & Vetter (2003)).

As a result, their synthesised faces looked very realistic, even in various lighting or viewpoint conditions. However, in practice, this method had several shortcomings, such as manual initialisation in the model-fitting process and problems with face reconstruction with partial face occlusion and severe lighting conditions.

A 3D method that also handled wide pose variations but did not require manual initialisation in the model-fitting process was presented by (Ding et al., 2016) (see **Figure 13**). This 3D method estimated the pose and shape of the face using only one 2D image. It detected a contour of the face and positions of five feature points (i.e. eyes, nose and corners of a mouth) to know their exact locations, where to transfer 2D face image in the 3D model and how to properly align a 2D face image onto the 3D face model.

After transferring the texture information, a pose normalisation technique was applied to the image to correct potential deformation that appeared in the facial texture due to the pose variations. Unfortunately, similar to the previous 3D method by Blanz et al. (2003), this method also could not cope with the problem of restoring the facial texture information that was lost due to occlusion. Thus, this method used only non-occluded facial textures to perform face reconstruction (see **Figure 13**), where each patch represented a specific feature of the face, which later was stored in PCA space. Storing data in PCA space helped to reduce data dimensionality and to lower the noise.

Due to the fact that the approach used limited information to estimate the pose and shape parameters of a face, generated reconstructions of a face were unprecise. Also, the method handled poorly face reconstruction in wide poses that fell into the range from -90 to +90 degrees in the horizontal plane.

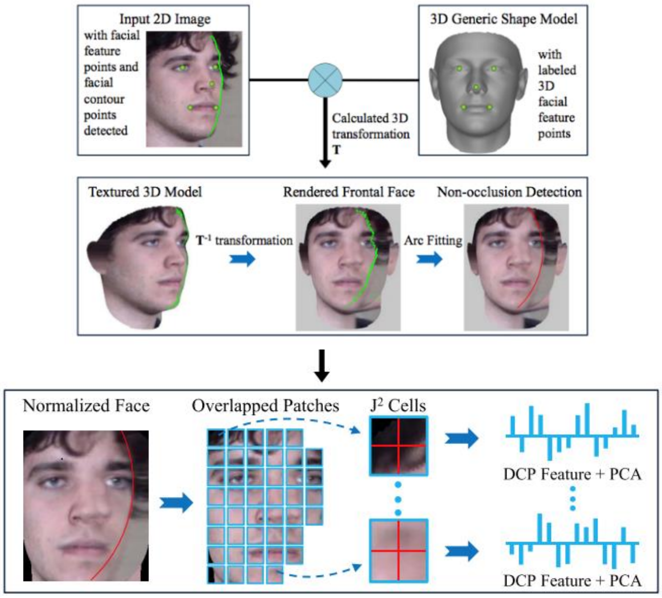


Figure 13 Face reconstruction method using only one 2D example view of a face (image reproduced from Ding et al. (2016)).

An approach that was better at handling face reconstruction in various poses of such a wide range was proposed by Jackson et al. (2017). They produced a 3D volumetric representation of the facial geometry and made spatial predictions at a voxel level. Their approach used a Convolutional Neural Network (CNN) as a database, which was trained on a dataset consisting of more than 60.000 2D facial images and 3D facial models. Jackson and his colleagues claimed that their method is capable of realistically generating facial images from faces presented in a wide range of poses, in various facial expressions or faces that are partially occluded. However, it was not capable of capturing fine details of a face (e.g. wrinkles or spots). Also, the approach required predetermine positions of the facial landmarks; otherwise, it poorly reconstructed facial expressions.

This was a brief overview of the main strategies for encoding and animating faces in computer graphics and their weaknesses. We will now proceed to introduce the key methods used in this thesis that addressed these weaknesses.

2.2. 2D view-based key methods

The 2D image-based multi-view method (Beymer & Poggio, 1995) or 3D face models such as polygonal meshes or image-based morphable face models (Blanz & Vetter, 2003; DeCarlo et al., 1998; Jackson et al., 2017; Waters, 1987), all these methods in their face reconstruction approaches took into account natural processes that take place in the human visual system how to achieve object constancy for faces. These methods required the implementation of sophisticated pre-processing and calculations and high-performance hardware where these calculations could be performed. However, none of them could obtain a reconstruction of a face fully invariant to wide poses and lighting changes. Therefore, this thesis introduces an alternative model that could replace complex face models and be able to realistically reconstruct faces in various viewpoints and illuminances, while using potential mechanisms built-in in the human visual system. This subchapter will describe key methods of this alternative model.

2.2.1. The biological inspiration for using PCA

In Chapter 1, a range of theories was presented that showed how face perception is performed in human vision. Faces are different in shape and sizes and contain many muscle tissues, bringing substantial complexity and diversity to a human face (Bruce & Young, 1998). Despite such a complicated structure of faces, people can perform face matching tasks without a lot of effort. In 1991, Valentine presented a framework called face-space to explain this human ability, which he insists is a potential psychological model that explains how the cognitive representations of faces are processed and stored in the human brain. The face-space framework made a great starting point for this study that aimed to explain how face recognition in humans is performed. This framework describes faces by multiple dimensions that encode the specific features of a face (e.g. internal, external or distinguishing) together with the space-based distance between the faces. Internal features can be the distance between the eyes; external features can be the length of hair, shape or size of a face, and even masculinity of a face; distinguishing features can be scars, tattoo or piercing. In this face-framework every individual face has its own position, depending on its dimensions. When the distance between the faces is categorised as similar faces, similar faces are placed closer to each other while different faces lay further apart.

Many researchers investigated the structure and dimensionality of a face-space framework, however, they were not able to come to one uniform theory that would explain how human vision is processing and storing facial information in the brain. One theory is that faces are distributed as norms (Giese & Leopold, 2005; Valentine, 1991). The norm-based model encodes faces based on their deviation from an average face located at the centre of a face-space. In this model, faces are represented by the vectors from an average face. The length of a vector shows the distinctiveness, and the direction of a vector shows features of a face or identity (Valentine et al., 2016). Another theory considers that faces are stored as exemplars (Lewis, 2004; Valentine, 1991). The exemplar-based model encodes faces as single points in the space without any specific reference to an average face. The distance between faces defines their measure of similarity to other faces; hence such distribution of faces within the face-space will point out the distinctiveness (Valentine, 1991). Similar faces will be located closer to the centre of the distribution for particular facial exemplar, while the distinctive faces will be located in areas of a low density. The debate is still ongoing. Nevertheless, the face-space framework became a theoretical structure for many empirical studies, which were combined with computational studies to investigate face recognition in human vision (Natu & O'Toole, 2011; O'Toole et al., 1993).

The face-space framework in this thesis was used as a potential structure of how the human visual system may store faces in the brain. This thesis used this framework in combinations with the psychophysical evidence presented in Chapter 1, which suggested that human vision processes face holistically and in a 2D manner rather than based on the sole local features of a face, and incorporated them into the Principal Component Analysis (PCA) technique to perform as a face-space to imitate the human visual system. The PCA approach in this thesis was chosen for a number of reasons as it works on holistic information of faces while holding face information in 2D, reduces the dimensions of big data sets, extracts and visualises essential visual information in faces that are required for face matching in various viewing conditions. The PCA component analysis method became a very popular tool in many research studies for its simplicity and successful performance in automatic face recognition (Blanz & Vetter, 2003; Blanz & Vetter, 1999). A detailed description of PCA space is presented below.

2.2.2. What is PCA, and what are eigenfaces?

Principal Component Analysis (PCA) is a statistical technique that reforms data from a high-dimensional (multidimensional) space to a low-dimensional space while retaining most of the data information. Karl Pearson invented the PCA technique in 1901 (Pearson, 1901). PCA uses linear transformations to find directions of axes, along which samples of a high dimensional dataset vary the most. This approach helps to keep samples with the most information while reducing the dimensionality of a dataset and improve the computational efficiency of the algorithm. PCA can be calculated with eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, but only after pre-processing a dataset, such as finding a mean vector of all samples and mean centring for every sample vector. Then from the calculated matrix, PCA will extract two separate components: directions of axes and variance. Axes that point to the most deviation of data are called principal components or eigenvectors. And variance that shows the value by which the eigenvector is scaled is called an eigenvalue. Sirovich and Kirby developed this approach further and tested PCA on a dataset composed of 2D face images (Sirovich et al., 1987). In their approach, they transformed every 2D face image from a database into a long vector. With PCA, they found principal components for that face database. These principal components are called eigenfaces. A small database of 9 images of faces was created to present the performance in a PCA space. Face images were taken from 5 different people in various expressions. Images were aligned based on a position between the eyes. Finally, these images were transferred into PCA space and were described with eigenfaces and eigenvalues (see **Figure 14**).

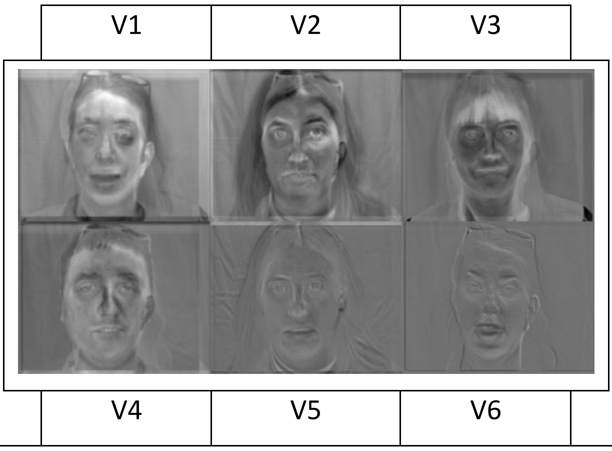


Figure 14 depicts six eigenfaces from PCA space, with maximum eigenvalue at V1 that becomes smaller when getting closer to V6.

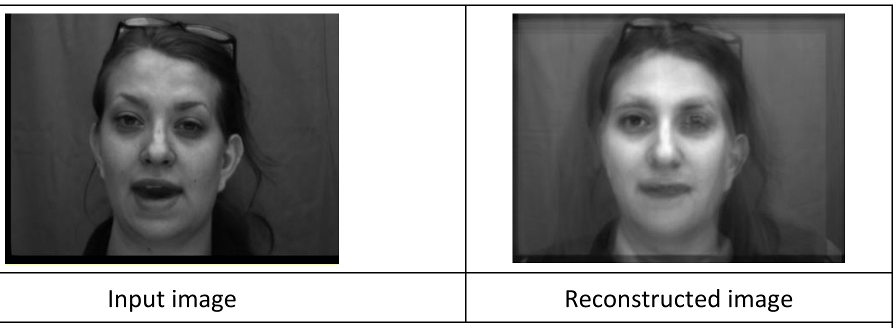


Figure 15 Input image reconstructed using only information of six eigenfaces.

Figure 15 shows the face reconstruction results of an input image taken from a training set. It can be noticed that facial reconstruction is relatively weak. However, we can observe reconstruction of glasses on a top of a head, features of a face and expression that mimics facial expression in the input image. Such weak reconstruction can be improved with better image alignment and with increasing the number of eigenfaces, but that would require to increase the number of facial images in the database. With an increased number of eigenfaces, the reconstruction of a face will look more like an input image. Sirovich and Kirby noted such results in their work (Sirovich et al., 1987) together with Cowe (Cowe, 2003) and Berisha (Berisha, 2009). According to Cowe, with an extensive database, face reconstructions can be made even outside of a training set.

PCA technique became a widespread approach to explain facial recognition. O'Toole et al. in 1993 used static faces to define face recognition for unfamiliar and familiar faces (O'Toole et al., 1993). They found that PCA encodes general information that is related to face recognition for unfamiliar faces in higher components, and later components encode unique information that is related to face recognition for familiar faces. Calder et al. in 2001 found that PCA components can encode facial identity and expressions (Calder et al., 2001). They used Ekman's and Friesen facial dataset (Ekman &

Friesen, 1976) to analyse facial expressions. This facial dataset consisted of various images of people with different facial expressions. In their work, PCA was successful in the recognition of facial expressions. Also, they discovered components separation, where components were tending to code just facial identity or facial expression. Principal components with a more significant variance were coding facial identity while components with smaller variance coded facial expressions.

Valentin showed that the PCA technique works like an auto-associative memory when Kohonen proved that this is how the human brain works— by finding that cortical networks are auto-associative where an element can retrieve a whole memory network from only a tiny sample of itself (Kohonen, 1986; Valentin et al., 1994).

Taking into account all the above mentioned, this thesis will incorporate a PCA technique to find eigenfaces from a facial dataset and to work as a content addressable memory that will be used to recover information stored in the PCA space. This technique in more details will be discussed in Chapter 3.

2.2.3. Facial mimicry tracking and optical flow

As we know, human skin has a very complex structure, and this raises significant difficulties in obtaining realistic reconstructions of a face for many methods that were discussed previously. In this thesis, an optic flow method will be used to track facial movements and transfer them onto a computer-generated face model. This model became very popular in the analysis of facial expressions in recent years (Beymer & Poggio, 1995; Burriss et al., 2007; Dimberg et al., 2002; Essa & Pentland, 1997). Optical flow technique presents the magnitude and direction of a facial movement. It became a handy method for tracking relatively small facial motions in video sequences because optical flow tracks motion between two 2D image frames at a pixel level.

Optical flow method can be described not only with computer vision but also with psychology. Psychology explains optical flow as an apparent motion of objects, or edges of a scene, obtained as a result of the movement of the observer (e.g., retinal image from the eye) relative to the scene.

An optic flow algorithm that will be used in this thesis is the Multi-channel Gradient Model (McGM). This model is a gradient-based optic flow technique that was developed by Johnston in 1999 (Johnston et al., 1999). It had been biologically motivated to introduce how motion perception processes may be accomplished in the human visual system. This model worked with spatiotemporal linear filters; the behaviour of simple cells in V1 motivated such decision. This model incorporated three spatiotemporal linear filters, which decreased its mathematical complexity. The first step of this model was to achieve a blurring effect in images by implementing convolution with Gaussian derivatives of various orders. Next, these blurred images were approximated with Taylor expansion up to first order. This step allowed to approximate and adjust image brightness at specific points in space and time. From these basic measures, Johnston and colleagues computed a matrix of the inner product. Finally, from this matrix, they found ratio values that presented a speed direction of every point in every image. The McGM model will be used in tandem with a view-based approach introduced by Cowe (Cowe, 2003) that will be described in more details in Chapter 3.

2.3. Chapter 2 summary

The human visual system is capable of easily defining people faces in various complicated and uncontrolled viewing conditions (e.g. in large viewing angle, poor illuminance, familiar or unfamiliar faces). Such capability is mounted in the human brain and works like a computer that runs vision algorithms. The psychophysical pieces of evidence introduced in Chapter 1, state that the human brain process faces holistically and in a 2D manner. The computational model of a face that will be built in this thesis is based on these biologically motivated processes involved in face reconstruction in various viewing conditions. To build a good computational model of the face is not an easy task. Hence, this chapter introduced the main strategies for encoding and animating faces in computer graphics addressing their weaknesses and how they relate to biological pieces of evidence on object constancy. One of the earliest strategies that were used to replicate and realistically reconstruct the complex structure of a face was the polygon method. First attempts to use the polygons method produced poor results, but it has been drastically improved in recent years to give spectacular results in realistic animated face reconstruction. However, this method required manual landmarks on a face to define which regions will need to be driven and integrated very complex mathematical computations with high-performance hardware to render these facial animations. Similar results and algorithm performance for face reconstruction was obtained with other image-based techniques. To address their weaknesses, several alternative models were developed using key methods based on natural processes that are happening in the human brain. One of the key methods is Principal Component Analysis (PCA). The PCA was integrated as a face-space framework, which was created to explain how face recognition in humans is performed. It is a well-known statistical technique that, while working on holistic information of faces, retains face information in 2D. The PCA component analysis method is a trendy tool in various research studies for its simplicity and successful performance in automatic face recognition. Another key method, that is biologically motivated and introduces how motion perception is processed in human vision, called Multi-channel Gradient Model (McGM). It is an optic flow algorithm that can be used to track relatively small facial motion in video sequences. This chapter closes the theoretical part of this thesis that describes its psychological and biological motivations on how the human visual system might work. Next chapters of this thesis will introduce empirical work which will investigate face constancy across varying viewing conditions, like object orientation, lighting and asynchronous data sets.

3. PCA-based computer-generated facial mimicry model

The main developments in the thesis are built on a PCA-based appearance model approach to encoding and representing dynamic faces. This chapter will introduce the general methods that were used in the thesis, including synchronised video capture from multiple views and correction for geometrical distortion. Then it will proceed to explain the approach that was used to solve the face constancy problem in different lighting conditions. Finally, the method used for generating example-based models of the face will be described[1].

3.1. Project's hardware and software specifications

This subsection will introduce a detailed explanation of the infrastructure and fundamental methods used in this thesis.

3.1.1. Camera capture

One of the significant technical challenges of this thesis was synchronously capturing multiple views of a face in non-rigid motion. In order to overcome this problem, both hardware that captures facial movement and software that performs calibration and PCA analysis had to be appropriately configured. For this purpose, a capture rig consisting of six moveable cameras was constructed, and cameras were mounted on the rigid arc (see **Figure 16**).

[1] The method was initially developed by Glyn Cowe (Cowe, 2003).

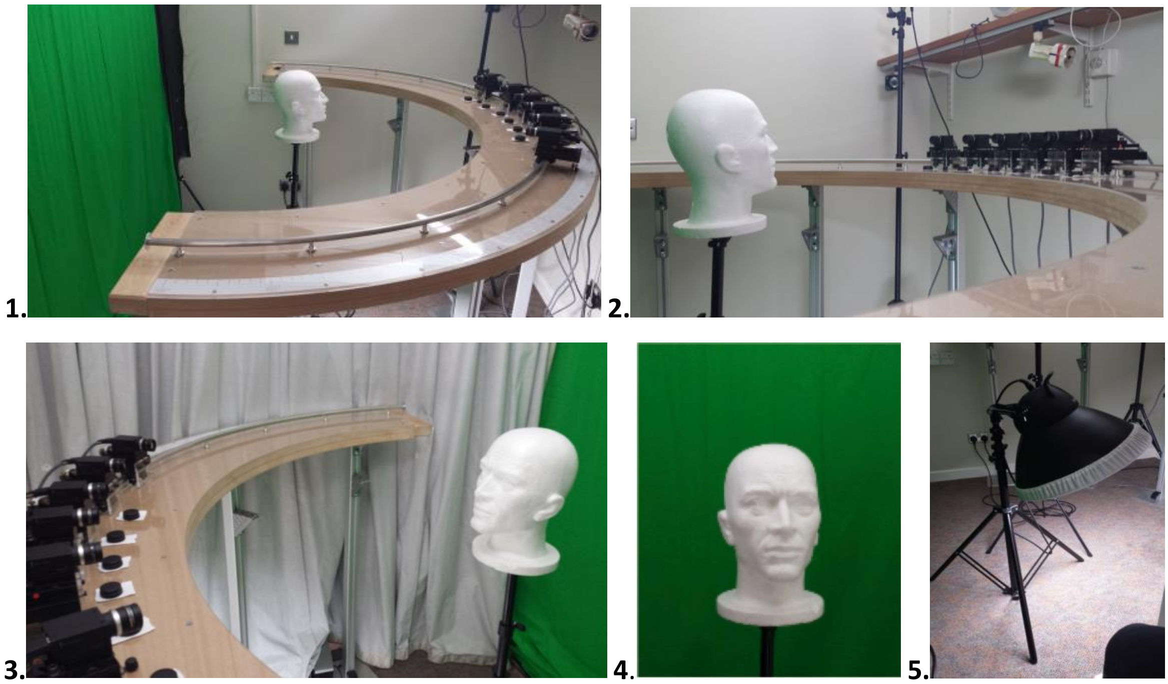


Figure 16 Images **1,2,3** represent a capture rig that consists of six cameras mounted on an arc construction. Image **4** represents the subject on which experiments were done. Image **5** shows the professional photography lamp on a stand which is placed in a way that the luminance conditions would be appropriate to reveal the most valuable information in the image.

To enhance details in the face, professional photographic lamps were installed on stands around the capture rig. Selecting a proper light level over the images is very important, as too dark or too bright images will exclude valuable information from the face pattern. Hence, luminance should be chosen optimally, and the illuminance should be distributed symmetrically on the subject's face (see **Figure 17**).

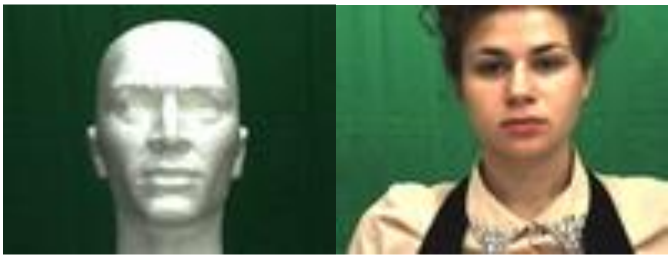


Figure 17 The plastic head (left), human face (right). The light source is symmetrically situated on the plastic head and has the same light adjustments, and a similar approach was applied to the human's face.

The participant (**Figure 16**, **Figure 17**) is located in the middle of the capture rig such that all six cameras can be focused on him or her. The camera images could be combined in pairs to recover the stereo disparity and binocular depth, although this was not implemented in this work. Two conditions were chosen for the camera separation. First was around 5 cm to simulate human vision, and the second was around 18 cm. A relatively high frame rate of 60 fps was used in order to capture fast and smooth non-rigid movements of the face (Brand, n.d.). The cameras (Grasshopper GRAS-03K2C (FireWire), PointGrey), which are used in this project (**Figure 18**), have high capture rates (60 fps) and resolution

(640x480). Also, all the cameras are synchronised across the 1394 bus on the same personal computer within 125 microseconds of each other. The cameras simultaneously deliver video streams at 640x480 image resolution with RGB 24 bit pixel format and 60 frames per second (fps) frame rate.

Figure 18 Camera (Grasshopper GRAS-03K2C (FireWire) produced by PointGrey) is used to capture the facial motions of the subject. The camera is mounted on arc formed capture rig. Adapted http://flir.com. Retrieved from URL https://www.flir.com/support/products/firewire-cameras/#Overview

Each camera has been updated and connected to a monitoring station. The monitoring station is a personal computer, Dell Intel(R) Xeon(R) CPU, with 12M Cache, 3.46GHz Clock Speed, 64bit Instruction set, 192GB of RAM and 500 GB of Hard Disk with the integrated operating system Windows 7 and development environment Microsoft Visual Studio 2010. Processes, synchronisation of the cameras data recording and storage in the hard drive of the monitoring station was accomplished in Microsoft VS 2010, using the C++ programming language with no lost images.

3.1.2. Camera calibration

Camera calibration procedures are a necessary step in the computer vision system in order to reconstruct points on an object surface. The synchronised cameras had the same settings (gain, exposure, sharpness, saturation, shutter, brightness). They also had the same angle and elevation parameters, which allowed here to obtain comparable information about the face from various angles. However, it was not enough to obtain aligned and precise data about the face. The obtained data contained distortions due to the camera lenses and misalignments of the vertical axis. The data were corrected in software using geometric methods. That simplified the processes of relating the different views of the face obtained from the multiple cameras

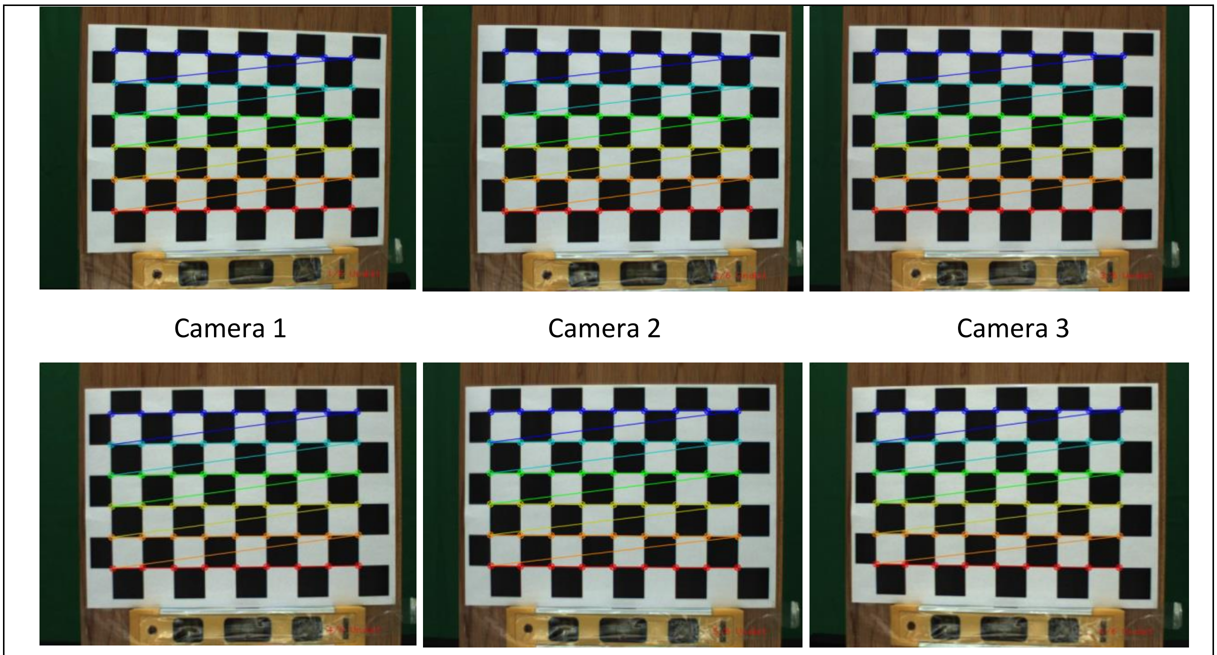
3.1.3. Camera settings

There are many methods of geometric camera calibration. These fall roughly into two categories *photogrammetric* and *self-calibration*. Photogrammetric calibration was introduced by Tsai (Tsai, 1987). It requires corresponding metric information from 2D images of the three-dimensional (3D) world. The method recovers the interior orientation, the exterior orientation, the power series coefficients for distortion and an image scale factor that best fit the measured image coordinates to known target point coordinates. The drawback of this method is that it is challenging to implement and store the objects. Another approach described by Zhang (Zhang, 2000) involves filming a 2D planar object (usually a chessboard plane) that is designed for calibration. This approach needs only six parameters to describe position and orientation. It is simpler, faster and cheaper than photometric calibration, as well as more precise and robust; hence this camera calibration approach has been integrated into the system.

A camera calibration procedure used in this work is modified from the original model of Zhang to allow for the six viewpoints and was accomplished in several steps. The software was made using the Visual Studio 2010 development platform, with the help of the OpenCV (Open Source Computer Vision) library and was written in the C++ programming language. After calibration, distortion coefficients were obtained, which were used to improve the camera's relation with real-world units. The steps in the procedure are in performed order:

***1.** Capture.* All cameras observe a reference object (planar pattern) in different orientations (in our case, six different viewpoints). The increased number of orientations (six, increased from two in the method of Zhang) improves the accuracy of the camera calibration.

***2.** Feature points detection in the images.* Take, for example, a planar pattern of a 9x6 chessboard. The simple geometry of the chessboard structure allows the features to be found easily. Because of these advantages, at present, most camera calibration procedures use 2D reference objects such as these classical black-and-white chessboards.



Camera 1 Camera 2 Camera 3

Camera 4 Camera 5 Camera 6

Figure 19 Camera calibrations that determine the camera's relation between the camera's natural units (pixels) and the real-world units (e.g. millimetres).

We can observe a significant lens distortion in the images collected from the cameras. The corners of each square are detected as the intersection of the straight lines fitted to each chessboard square (see **Figure 19**).

***3.** Estimate the camera projection matrix.* Five intrinsic parameters, as well as all the extrinsic parameters, should be estimated in order to obtain the camera projection matrix, which describes parameters that will be used to transfer 3D points from world coordinates into 2D image points

A 2D point position in pixel coordinates can be denoted as $[u\ v\ 1]^T$. A 3D point position in world coordinates is introduced as $[x_w\ y_w\ z_w\ 1]^T$. The relation of 2D points in pixel coordinates with 3D points in the world coordinates is represented in Formula (3.1)

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = A[R \quad T] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{3.1}$$

Here A is defined as the *intrinsic matrix* (see Formula 3.2)

$$\mathrm{A} = \begin{bmatrix} \alpha_x & \gamma & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.2}$$

The intrinsic matrix contains five parameters. The parameters are the *focal length* (measures amount of light that is captured by the lens), the *image sensor format* (which determines the angle of view of a specific lens with a specific camera), and the *principal point* (which is the point where the principal plane intersects the axis). The parameters $\alpha_x = f * \mathrm{m}_x$,$\alpha_y = f * \mathrm{m}_y$ represent the focal length in terms of pixels, where m_x *and* m_y are scale factors relating to distance, and f is the focal length in terms of distance. The γ (Hartley & Zisserman, 2003) parameter describes the coefficient of skewness of the two images axes *x* and *y* (often equal 0). u_0 *and* v_0 are the optical centres, which would ideally be in the centre of the image, expressed in pixel coordinates (**Figure 20**). The *R, T* are parameters of an extrinsic matrix; they define how the system will transform from a 3D external world coordinate system into a camera coordinate system. T is a translation vector that will show a new origin position in the camera coordinate system. The *R* is the rotation matrix that is used to encode camera orientation with respect to a given world frame.

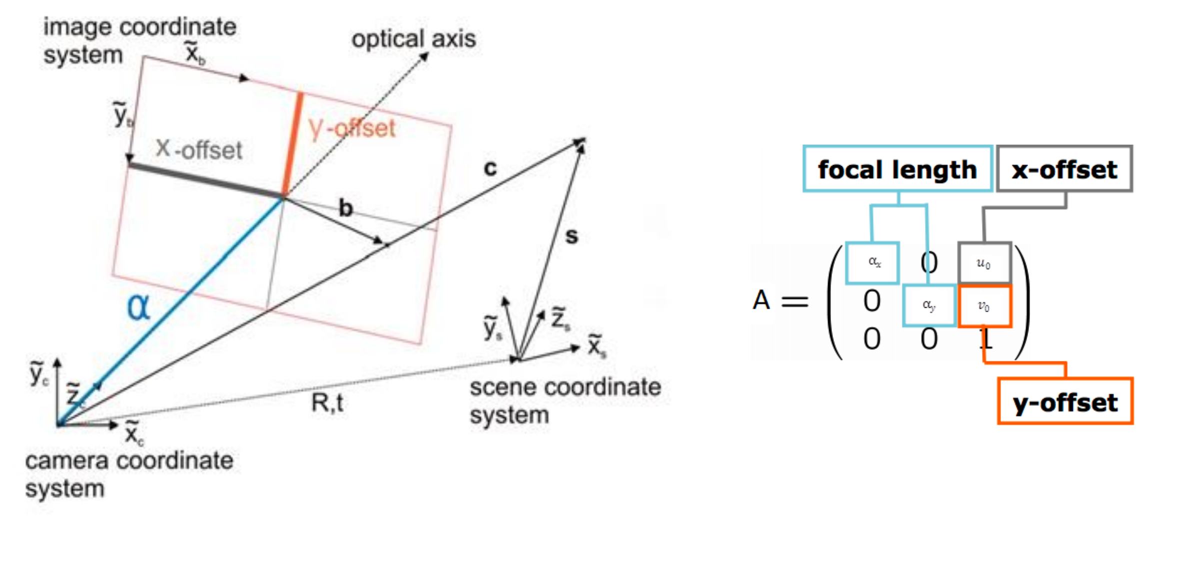


Figure 20 shows what parameters of a camera are used to build the intrinsic and extrinsic matrices. Image adapted from ais.informatic.uni-freiburg.de. Reprinted 17 February 2020, page 12, from http://ais.informatik.uni-freiburg.de/teaching/ws09/robotics2/pdfs/rob2-08-camera-calibration.pdf.

4. Calculate the camera's lens distortion. Lens distortion has two nonlinear effects: Radial and Tangential distortions. These distortions occur when physical elements in a lens are not perfectly aligned (**Figure 21**).

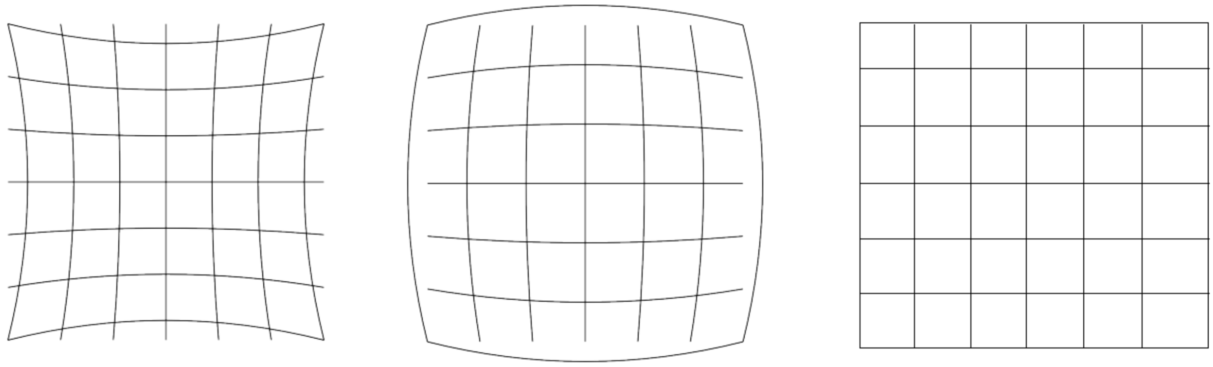


Figure 21 Negative radial distortion "pincushion" (left). Positive radial distortion "barrel" (middle). No distortion (right).

To obtain a corrected image point, the following set of formulae was used:

1) Project $(u\ v\ z)$ to "normalised" image coordinates (Formula 3.3).

$$x' = x/z$$

$$y' = y/z \tag{3.3}$$

where x', y' are undistorted image point as projected by an ideal pin-hole camera.

2) Apply the radial and tangential distortion (Formula 3.4).

$$r^2 = x'^2 + y'^2$$

$$x'' = x'(1 + k_1 r^2 + k_2 r^4) + 2p_1 x'y' + p_2(r^2 + 2x'^2)$$

$$y'' = y'(1 + k_1 r^2 + k_2 r^4) + p_1(r^2 + 2y'^2) + 2p_2 x'y' \qquad (3.4)$$

where k_1, k_2 are the radial distortion coefficients and p_1, p_2 are the tangential distortion coefficients.

3) Apply the focal length translation of the image centre (Formula 3.5),

$$u = \alpha_x * x'' + u_0$$

$$v = \alpha_y * x'' + v_0 \quad , \qquad (3.5)$$

where u, v are 2D points expressed in pixel coordinates.

The camera calibration algorithm was applied to images. Results are presented in **Figure 22**.

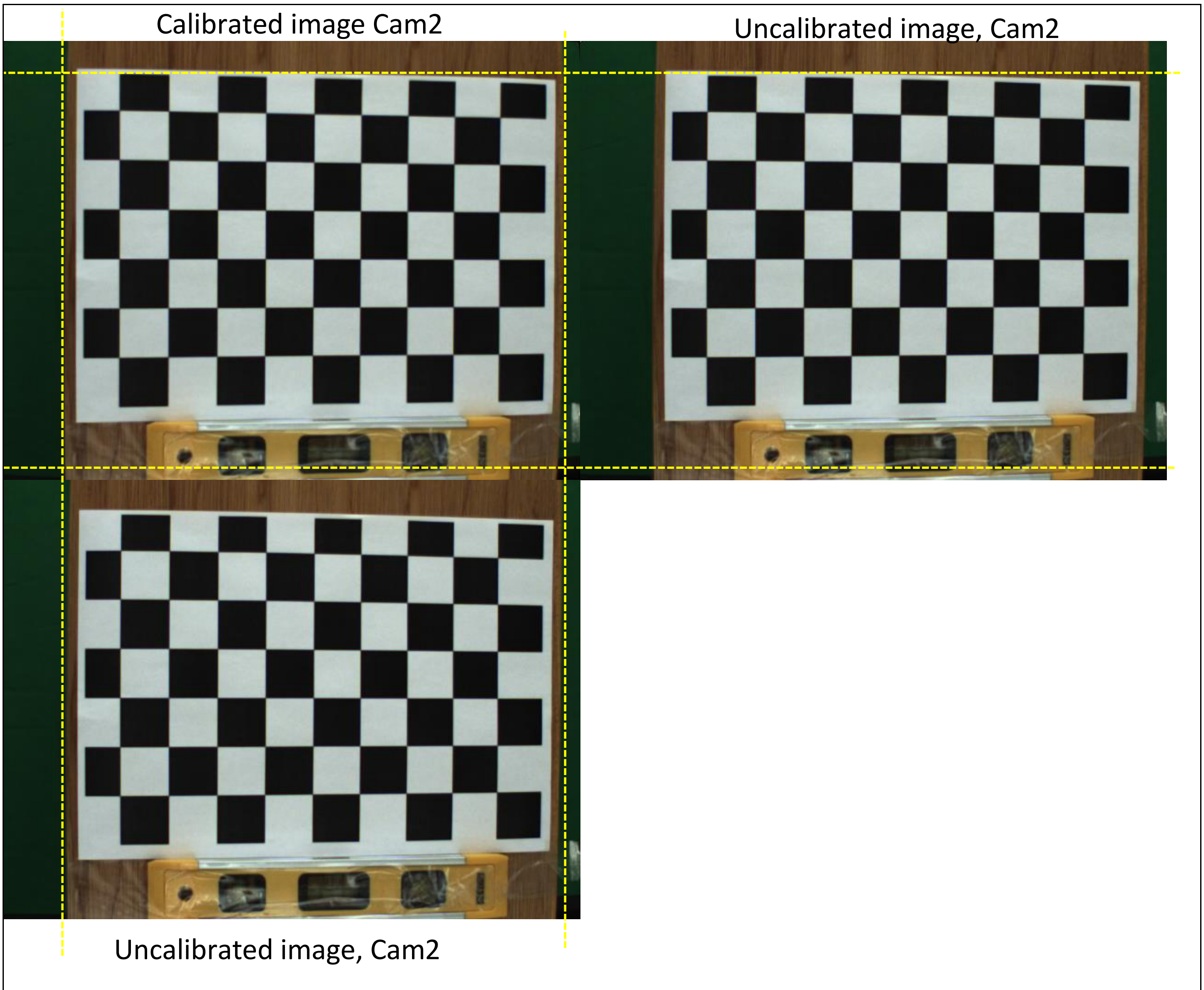


Figure 22 Results of applying the camera calibration algorithm to images. The yellow dashed line represents changes after calibration in images along the horizontal and vertical axis. Notice the difference in the position of the checkerboard between the images. For instance, compare the calibrated image (top left) with the uncalibrated image in the bottom left. Inspection of the dashed yellow line reveals a gap near the bottom right of the checkerboard that is larger in the uncalibrated image than in the calibrated image. Next, compare the calibrated image (top left) with the uncalibrated image in the top right. Again, the dashed yellow line reveals a gap near the top right of the checkerboard that is slightly bigger in the uncalibrated images that in the calibrated.

3.1.4. Geometric correction

The epipolar geometry is a geometry that explains the relation between the two views. This geometry helps search for corresponding points in the process of stereo matching when an alignment problem arises from different captured images. In this thesis, this problem appeared due to the misalignment between cameras. In particular, the position of the subject over the captured images was slightly shifted along the horizontal axis. That can easily be noticed by following a blue dashed line over the images in **Figure 23**.

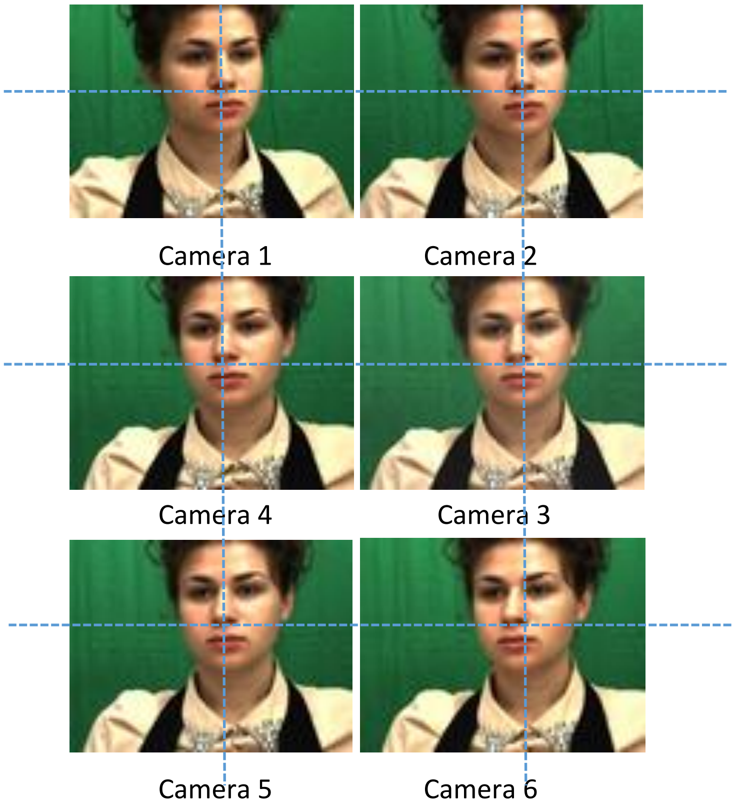


Figure 23 Alignment problem. The nose position changes between the cameras; it can easily be noticed by looking at the blue dash line over the images.

The subject’s position of the nose stays relatively constant along a vertical line, in contrast with changing nose position along the horizontal line.

To correct this misalignment, all images were aligned based on the nose bridge of the subject. Therefore, the nose bridge was located on the subject’s face in each image. Nose bridge detection was accomplished using Haar feature-based cascade classifiers introduced by Paul Viola and Michael Jones in 2001 (Viola & Jones, 2001). This classifier is a machine learning approach and, therefore, needs to be trained. In order to train this technique, various data was used, for example, positive (images of faces) and negative (images without faces) images, to obtain better detection of objects in other images. Although epipolar geometry arises from the studies of stereo vision, it is also helpful in solving the general alignment problem. From **Figure 24**, we can observe two cameras pointed at the same object but from different view perspectives. The objective of epipolar geometry is to describe the relation between the two cameras’ resulting views.

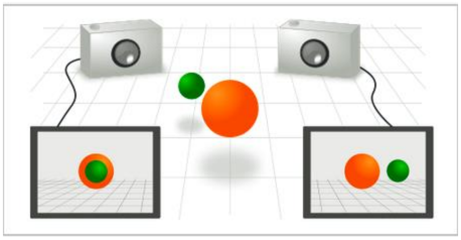


Figure 24 An example of epipolar geometry that will work on finding a relation between two different images of the same scene that taken simultaneously by two cameras, placed only in different perspectives. Reprinted from Wikipedia, 13 February 2008, by Author Arne Nordmann. https://en.m.wikipedia.org/wiki/File:Aufnahme_mit_zwei_Kameras.svg

To start describing an epipolar geometry, first, the three most crucial epipolar geometry constraints must be defined. These are the epipolar plane($\boldsymbol{C_L Q C_R}$), epipolar lines($\boldsymbol{e_L q_L}$ and $\boldsymbol{e_R q_R}$) and epipoles($\boldsymbol{e_L}$ *and* $\boldsymbol{e_R}$) (see **Figure 25**).

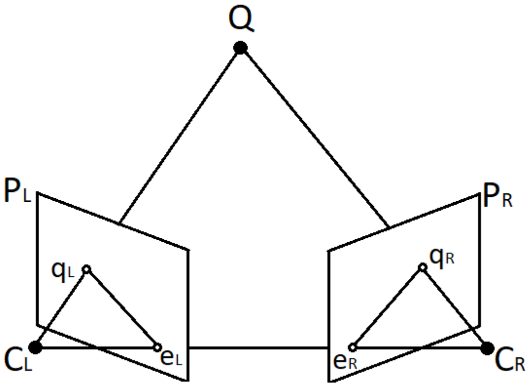


Figure 25 The epipolar constraints. Adapted from author ZooFari. Retrieved from URL https://commons.wikimedia.org/wiki/File:Epipolar_Geometry1.svg

In **Figure 25**, there are two cameras C_L, C_R and 3D point Q (Object point) that create an ***epipolar plane*** $C_L Q C_R$ that is placed between these points. Object point Q is projected in the point q_L in the 2D image plane of the left camera P_L also in the point q_R in the 2D image plane of the right camera P_R. The ***epipoles*** e_L *and* e_R are the points that intersect with baseline C_L *and* C_R and image plane P_L *and* P_R. The baseline is a joining line between the optical centres of the left and right camera C_L *and* C_R. The point q_L in the left camera image was created because of ray Qq_L. This ray is then projected onto the right camera's image plane, and it creates a line $e_R q_R$, that is called the ***epipolar line.*** The object representation from point Q on the image plane of the right camera will always lay on the epipolar line $e_R q_R$. Thus, each point q_L in the image plane of the left camera corresponds to an epipolar line $e_R q_R$ in the image plane of the right camera. In this case, a pair (corresponding point) for point q_L in the image of the right camera will lie only on the corresponding epipolar line $e_R q_R$. Similarly, each point q_R in the right image corresponds to an epipolar line $e_L q_L$ on the left.

So epipolar geometry searches for stereo pairs or verifies whether points of a pair build a stereo pair (i.e., a projection of some point in space). Epipolar geometry expressed in coordinates has a very straightforward notation. Assume we have a pair of calibrated cameras and let q_L be a point in the image of one camera written in homogenous coordinates and q_R be a point in the image of another camera also presented in homogenous coordinates. The relation of these points q_L, q_R can be found by a fundamental matrix $\boldsymbol{F}$ if these points satisfy relation in (3.6).

$$q_R{}^T F q_L = 0 \tag{3.6}$$

The matrix $\boldsymbol{F}$ is a fundamental matrix with a size of 3 x 3. Its rank is equal to 2, it is determined up to a nonzero factor and depends only on the matrices of the source cameras $P\ and\ P'$.

The equations of the epipolar lines are also calculated with the fundamental matrix. For the point q_L, the vector defining the epipolar line will have a form $e_R q_R = F q_L$, and the equation of the epipolar line itself will be $(e_R q_R)^T q_R = 0$. Similarly, for the point q_R, the vector defining epipolar line will have a form $(e_L q_L) = F^T q_R$. Also, rays that define projected points q_L and q_R onto the camera's image planes, need to be coplanar, i.e., lay on the same epipolar plane, to satisfy (3.6). It is a necessary condition that points would form a stereo pair.

To obtain the most reliable set of correspondences in images, the SIFT (Scale Invariant Feature Transform) algorithm was used proposed by David Lowe in 1999 (D.G. Lowe, 1999). This algorithm detects and describes local features in images at multiple scales and positions. These features are invariant to translation, rotation and re-scaling of the image. In addition, to find the best matches from detected features, we used the FLANN (Fast Library for Approximate Nearest Neighbour) algorithm introduced by Muja and Lowe (Muja & Lowe, 2009), which performs a fast search of nearest neighbours in a high dimensional space (**Figure 26**).

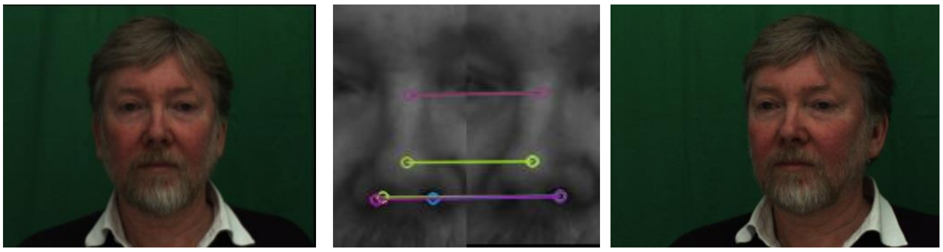


Figure 26 SIFT (taken from Lowe 1999) and FLANN (taken from Muja & Lowe 2009) methods practical results over the camera images by finding feature points and match them with other image feature points. The match of feature points is presented with colour lines (middle). Reference image (Camera 3 (left)), Unaligned image (Camera 5 (right)).

These two algorithms are very convenient for the calculation of the fundamental matrix. A robust statistics tool such as RANSAC (RANdom SAmple Consensus) (Fischler & Bolles, 1981) was used to estimate F (fundamental matrix) over the most massive possible set of correspondences obtained with SIFT FLANN. First, the RANSAC method was randomly selecting two points in the possible N points dataset obtained with SIFT. Next, it was calculating an error value between the estimated solution and the rest of the points. If the error value would appear to be less than some initially chosen threshold value, all the calculations instantly will need to be stoped, and the method had to go back to the first step and repeat the process from the begging. RANSAC repeated the calculation process until the best possible fit of a model was generated. In other words, the RANSAC method was used to calculate homography between two images and then eliminating weak feature pairs.

The process of computation of the fundamental matrix and model estimation with an 8-point algorithm looks as follows:

Fundamental matrix 3x3 is used to capture a relationship between corresponding points in two stereo views. In epipolar geometry, two corresponding points $x = (x, y, 1)$ and $x' = (x', y', 1)$ in homogenous coordinates of stereoviews (here (x,y) are the pixel coordinates) can be related with epipolar line. Epipolar line $Fx = 0$, shows where the corresponding point x' must lie in the other image. Hence, the formula of relationship for all corresponding pairs will look as follows (3.7)

$$x'^{T}Fx = 0 \tag{3.7}$$

This equation can be rewritten into (3.8)

$$Af = 0 \tag{3.8}$$

where A is matrix $n \times 9$, presented in (3.9) with $n = 8$, because of the chosen 8 point algorithm and $f = F_{11}, F_{12}, F_{13}, F_{21}, F_{22}, F_{23}, F_{31}, F_{32}, F_{33}$ is a column vector that stores values of fundamental matrix F.

$$A = \begin{pmatrix} x_1{}'x_1 & x_1{}'y_1 & x_1{}' & y_1{}'x_1 & y_1{}'y_1 & y_1{}' & x_1 & y_1 & 1 \\ x_2{}'x_2 & x_2{}'y_2 & x_2{}' & y_2{}'x_2 & y_2{}'y_2 & y_2{}' & x_2 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n{}'x_n & x_n{}'y_n & x_n{}' & y_n{}'x_n & y_n{}'y_n & y_n{}' & x_n & y_1 & 1 \end{pmatrix} \tag{3.9}$$

The easiest way to solve equation (3.9) is to find singular value decomposition (SVD) of A. Where $A = UDV^T$, where U and V are orthonormal matrices and D is a diagonal matrix containing singular values. The values of f are defined by components of a column V matrix that define the smallest singular value. Finding the F fundamental matrix from vector f can look like an easy task. However, if an image has noise, a task may become complicated. We know that the fundamental matrix has rank 2, but in this case, the rank of the fundamental matrix will be different. It means that epipoles and epipolar lines will either not show their real position or will be randomly scattered in the image. To solve this problem, we will find singular value decomposition (SVD) of F and decompose it into matrices $F = U \sum V^T$. Then, we set the smallest singular value at position (3,3) in the matrix $\sum to$ 0, this will give us a required rank of 2. Finally, the fundamental matrix can be easily calculated with $F = U \sum_2 V^T$

The result of this resolved alignment problem is represented in **Figure 27**. It can easily be seen that the nose position is identical across all images, despite differences in the camera position.

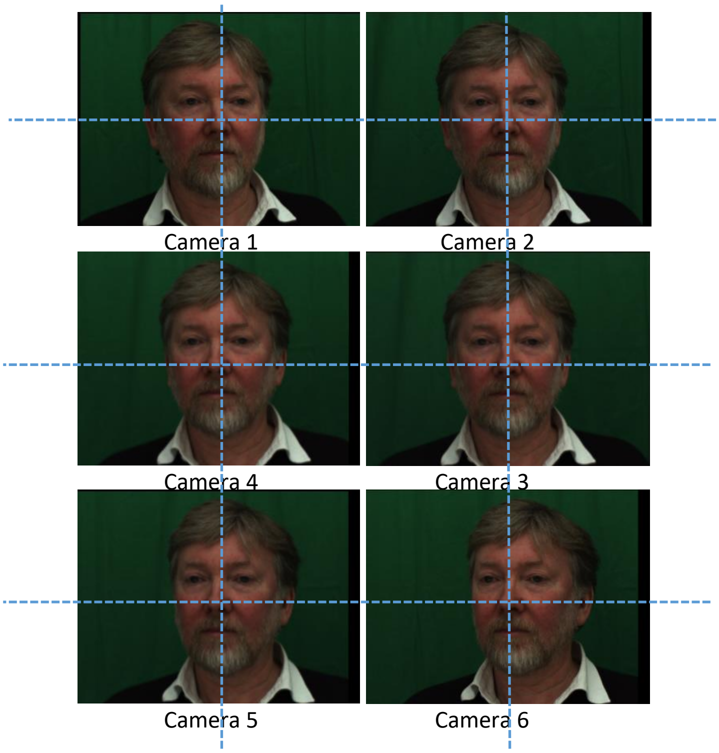


Figure 27 Alignment problem solved. The nose position stays relatively the same between the cameras; it can be easil noticed by looking at the blue dash line over the images

3.1.5. Lighting

This subchapter will explain how we set up the equipment to address the lighting invariance problem. To obtain synchronised non-rigid data of the face in various poses and different illuminance conditions is a very complicated task. To solve this task, we came up with the strategy of illuminating the face using primarily red, green and blue direct lighting from projectors in different locations. Then we separated each of the RGB colour channels from the facial images. The resulting grey-level images mimicked the effect of illuminating the faces with white light from different directions. First, we modified the capture rig by adding three projectors. These were QUMI Q5 projectors mounted on stands fixed to the capture rig, which also contained the six portable cameras mounted on a rigid arc (**Figure 28**). The three projectors were centred between each of the three pairs of cameras. Projectors were placed above the camera level to prevent direct interactions between projectors and cameras while collecting the non-rigid facial data.

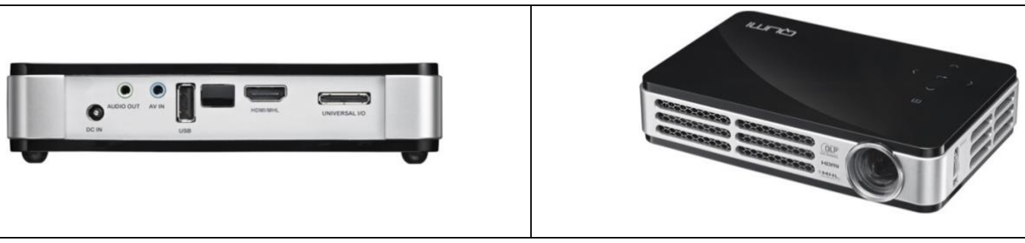


Figure 28 Qumi Q5 projector. Image adapted from vivitek.eu, n.d., Retrieved February 17, 2020, from https://www.vivitek.eu/Category/Discontinued-Projectors/3/Qumi-Q5

The QUMI Q5 projectors were chosen for their ability to produce a 60Hz frame-sequential signal. It proved beneficial to match the camera and projector's frame rate to reduce asynchronous flickering in the captured images. Moreover, the projectors could generate 500 ANSI Lumens, allowing good control over the brightness level, which was necessary for our experiment. Also, the projectors were portable and easy to install. The Q5 displayed controlled lighting by projecting an image (640x480) in native format (PNG) from a USB flash drive. Each projector controlled only one of the R, G, B colour lights (see **Figure 29**). The red light had (255.0.0) colour value, green – (0.255.0) colour value and blue – (0.0.255) colour value.

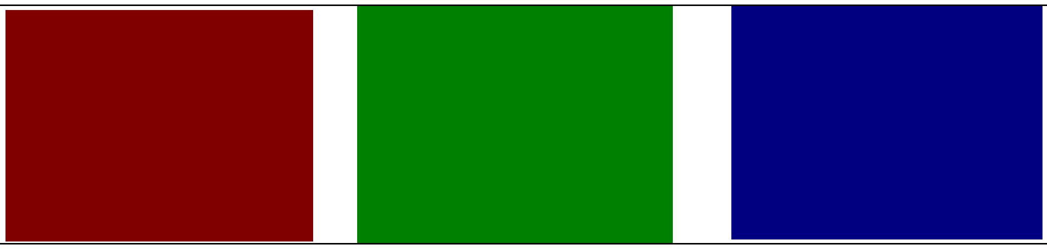


Figure 29 Colours that were used in projectors to control lighting. Each colour was saved as separate .png images with a resolution of 640x480.

The projectors with controlled lights were distributed in such a way that they would project light over the whole face (**Figure 30**).

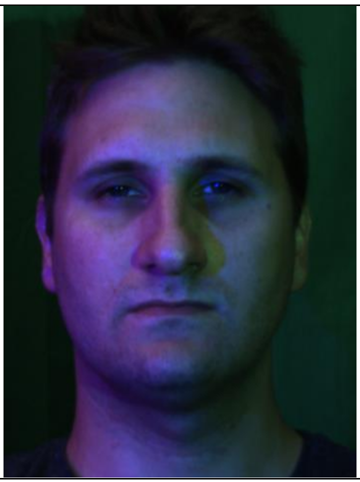


Figure 30 All three controlled lights from projectors are distributed to cover the whole human face.

Red controlled light covered the right side of the face, green covered the left, and blue colour covered the middle part of the participant's face (**Figure 30**). The position of controlled lights was set to obtain the illusion of different lights in different locations in the face. Applying only direct lights to the face images appeared to be having some dark areas (neck area, background, hair area). Hence, a new idea came up to add some diffuse light in addition to the direct lights; a diffuse white light was used to illuminate those dark areas in a face.

Moreover, in this experiment, white light was used in four different positions, white directional lighting with the same spectrum was placed in different places so that it could mimic the effect of the different lights in different locations. It is worth to mention that there was no change in direct controlled lights; the only change was the position of diffuse white lighting. Next, we have collected four data sequences by altering the position of an additional diffuse white illuminant while keeping the direct coloured lighting in the same position. The collected data sequences were in RGB 24 bit pixel format, where one 8 bit byte was allocated for each Red, Green, and Blue colour component. In each component, the value of 0 refers to no contribution of that colour, 255 refer to fully saturated contribution of that colour. Then we split the Red, Green and Blue channels of an RGB image into independent greyscale images. For maximum efficiency, we had to take into account that in an RGB image, the brightness is not derived equally from the three colour channels.

The green component has the highest gain with respect to the brightness of our image, red is less important, and blue is the least significant. The green channel holds the most details about the source image when the red channel has low contrast, and the blue channel collects more noise. Hence, the brightness level adjusted so that most of the facial features would be visible. **Figure 31**, **Figure 32**, **Figure 33**, **Figure 34** present four different illumination conditions, where only the position of diffuse white lighting was changed.

The first illuminance condition was obtained with a diffuse white light covering the left side of the participant's face (**Figure 31**). The second illuminance condition was obtained with the white light covered the right side of the participant's face (**Figure 32**). Third illuminance condition was created with diffuse light from below, highlighting the lower part of the participant's face (**Figure 33**). The fourth illuminance condition was uniform, with white diffuse illuminant distributed symmetrically over the participant's face (**Figure 34**).

The top row in **Figure 31** presents the first illuminance condition; here, we may observe a human face presented in ¾ view illuminated with red, green and blue lights with the white diffusive light positioned to highlight the left side of the face. The middle row presents three images obtained from splitting channels R, G and B channels of RGB image into separate greyscale images. The histograms in the bottom row show how pixels are distributed in each image for each different channel. These 8-bit grayscale images have 256 different grey values for each pixel. As it was mentioned earlier, we have increased the brightness value for all of the greyscale images in all illuminance conditions. The biggest brightness value was added to the face from the blue channel, the red channel had a small increase in image intensity, and the green channel was changed least. These brightness variations do not have an essential influence on further processing. This process of brightness rescaling was repeated for all the other data in the four different illuminance conditions. Please see **Figure 32**, **Figure 33** and **Figure 34**.

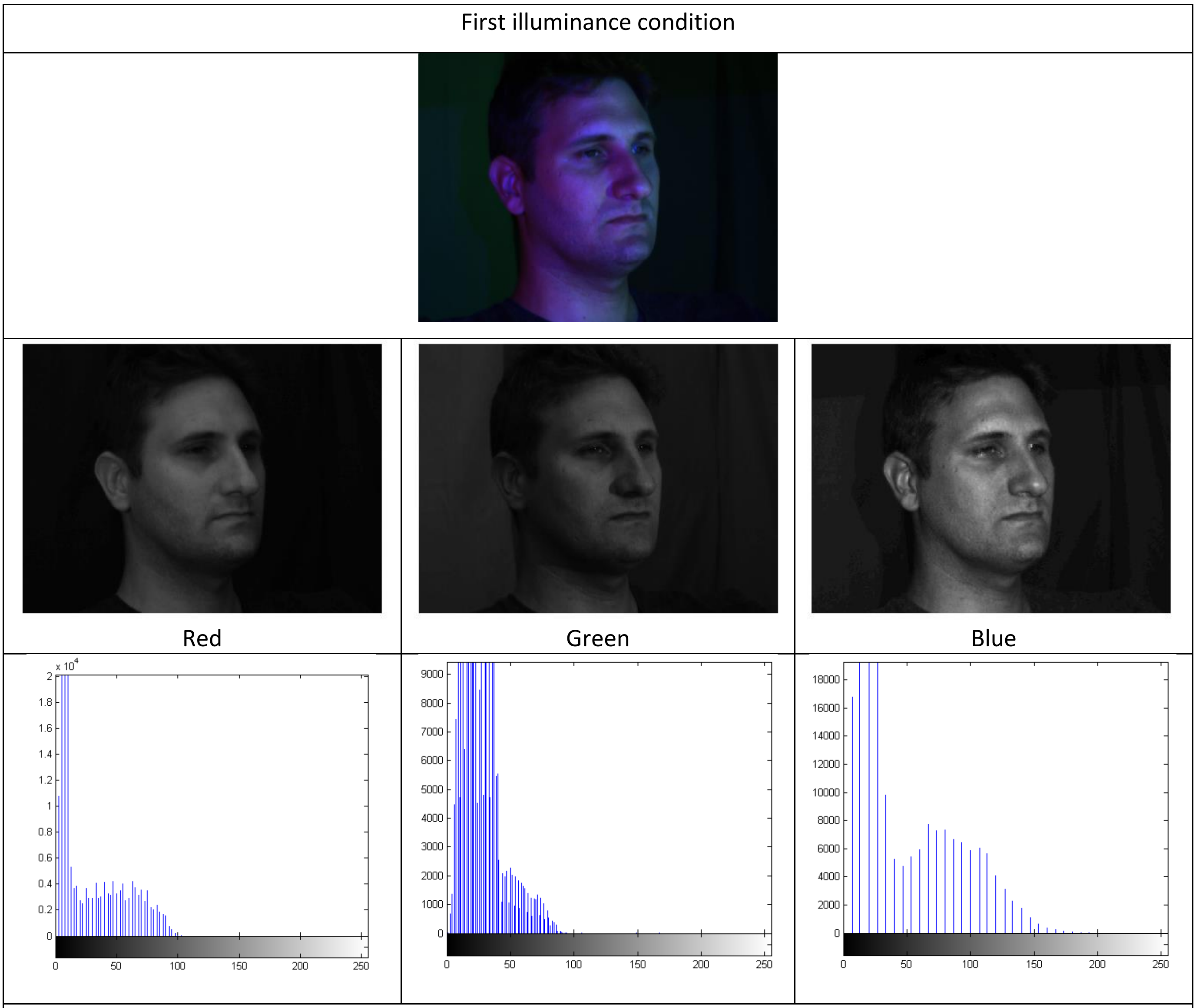


Figure 31 The figure presents the strategy of obtaining data for the illuminance experiment of light coming from the left side. The top row depicts an image that we will use to extract three colour channels. The middle row presents images of extracted colour channels. The lowermost row presents a histogram which shows a number of pixels and their intensities in every image from the middle row. We may observe that pixel intensity is much higher for a blue channel; for this reason, the blue image looks brighter than images captured in red and green channels, which contain lower pixel intensities.

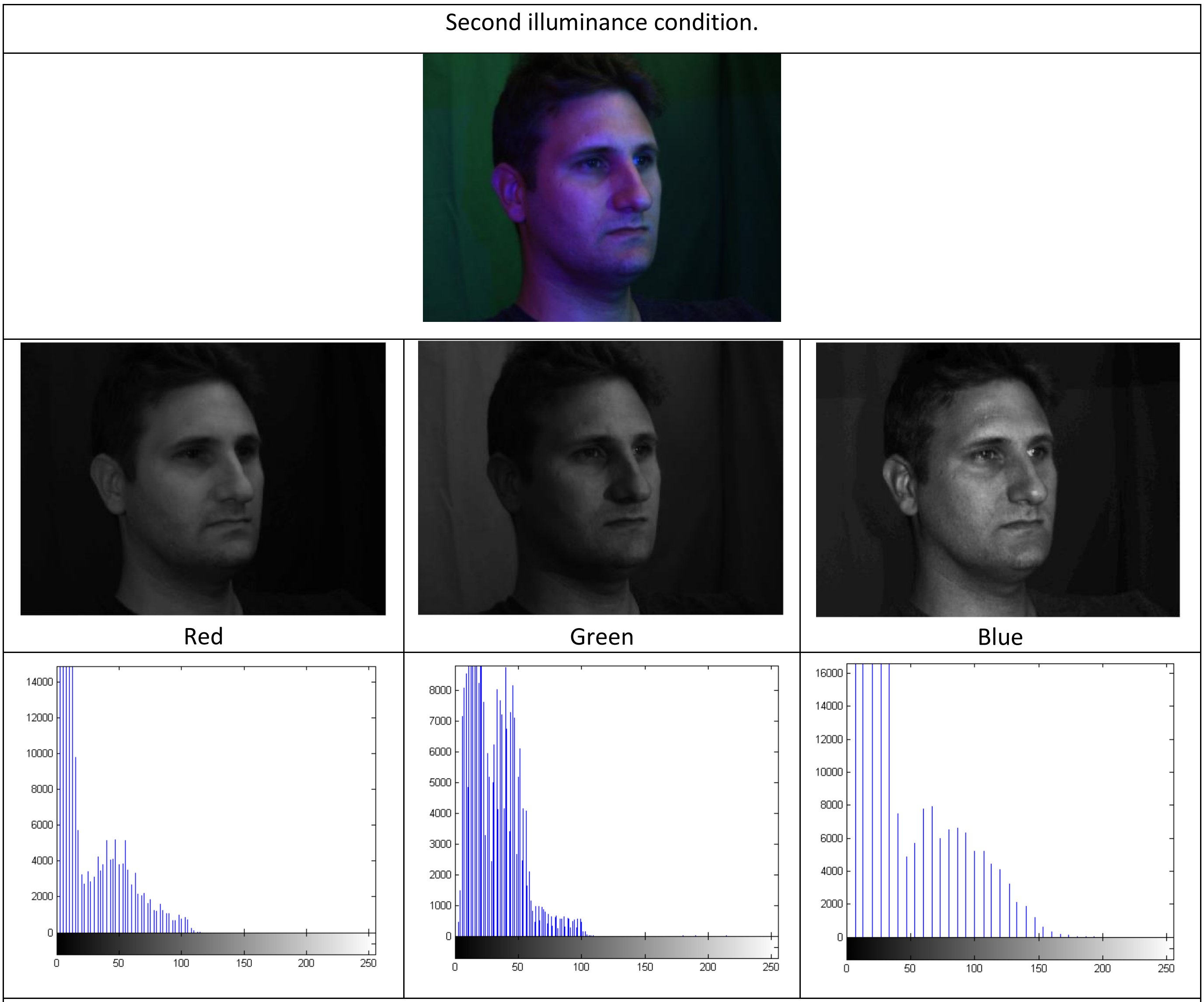
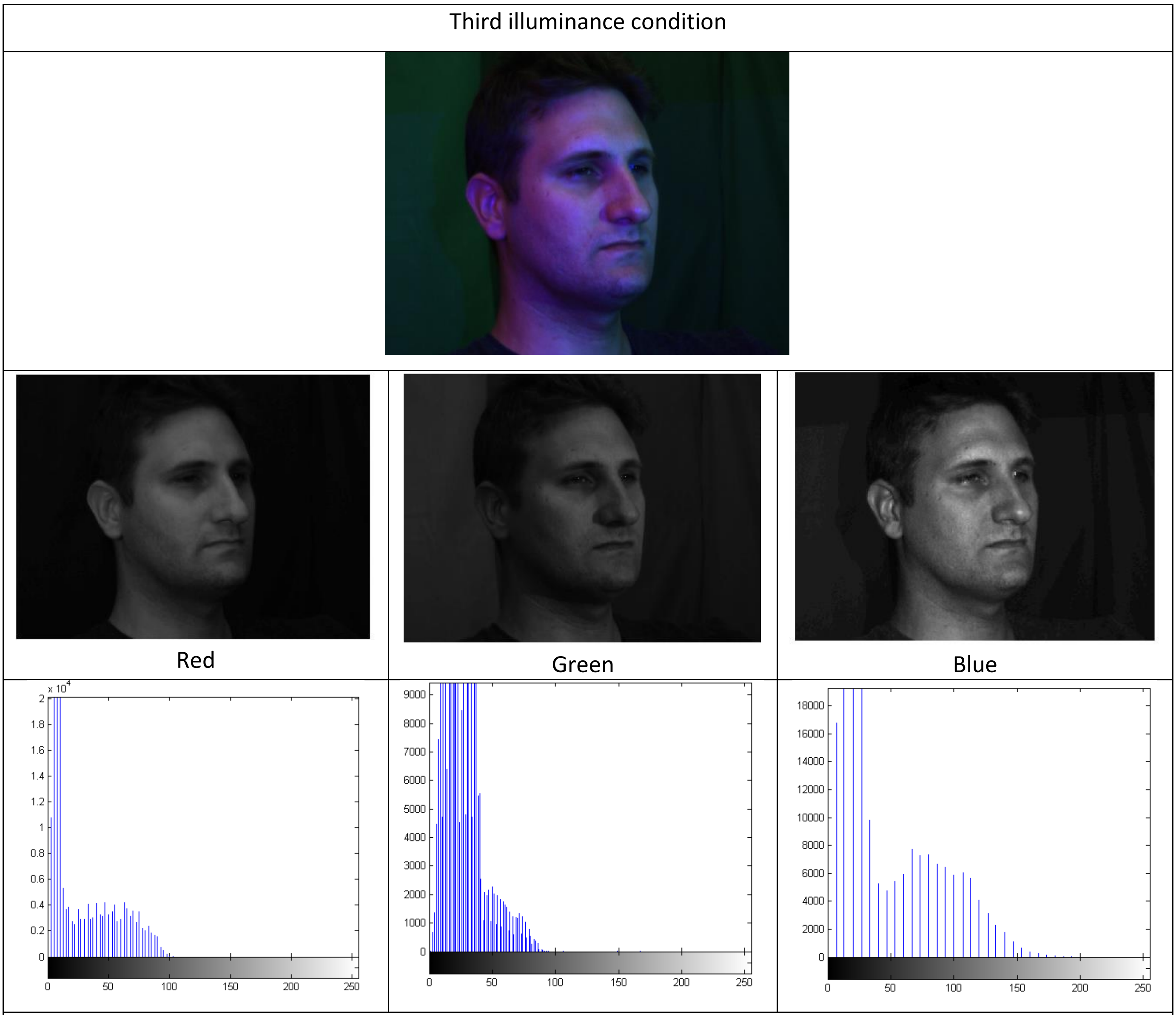


Figure 32 The figure presents the strategy of obtaining data when the light source is set on the right-hand side. The top row depicts an image that we will use to extract three colour channels. The middle row presents images of extracted colour channels. The lowermost row presents a histogram which shows a number of pixels and their intensities in every image from the middle row. We may observe that pixel intensity is much higher for blue channel, in this reason blue image looks brighter than images captured in red and green channels which contain lower pixel intensities



Third illuminance condition

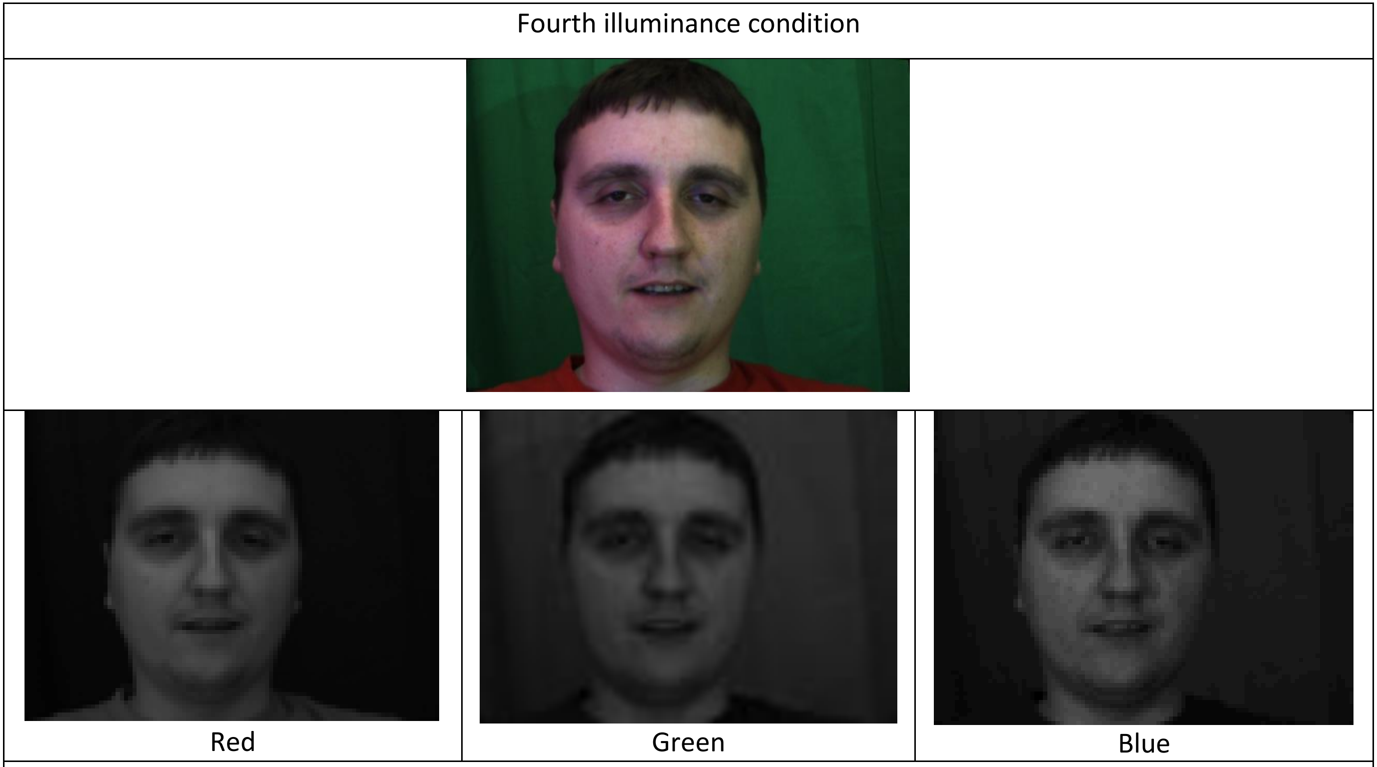
Red

Green

Blue

Figure 33 The figure presents the strategy of obtaining data when the light source is set at the bottom. The top row depicts an image that we will use to extract three colour channels. The middle row presents images of extracted colour channels. The lowermost row presents a histogram which shows a number of pixels and their intensities in every image from the middle row. We may observe that pixel intensity is much higher for a blue channel; for this reason, the blue image looks brighter than images captured in red and green channels, which contain lower pixel intensities.

The fourth light condition was a uniform, diffuse illuminant distributed symmetrically over the participant's face (**Figure 34**). This light condition differs from others regarding the intensity of controlled direct lights relative to the diffuse light. In **Figure 34,** it can be seen that the extracted colour channels have relatively low contrast and are less differentiated in relation to the lighting direction.



Fourth illuminance condition

Red

Green

Blue

Figure 34 The figure presents the strategy of obtaining data when uniform light distributed symmetrically over the human's face. The top row depicts an image that we will use to extract three colour channels. The middle row presents images of extracted colour channels. We may notice that due to uniform light, the effects of controlled direct lights are muted over the human face. This procedure gave us small changes in illuminance over that human's face.

The reason we chose to develop this method to solve a lighting problem is that now we have a set up that delivers three different illuminants for every single frame. There is no need to switch between illuminants on a frame by frame basis, which would be very complicated to accomplish. This method is straightforward. The methodology of this infrastructure will be described in more detail in chapter 5.

3.2. PCA method.

Principal component analysis is a multidimensional statistical analysis that is commonly used to reduce data dimensions with a minimal loss of useful information (Pearson, 1901). From a mathematical perspective, the PCA method is an orthogonal linear transformation that maps data from the original feature space to a new space of a lower dimension.

In this case, the first axis of the new coordinate system is constructed in such a way that the data dispersion along it would be maximised. The second axis is placed orthogonally to the first axis so that the variance of the data along it would also be maximised relatively to their remaining possible ones and so on. The first axis is called the first principal component, the second axis - the second principal component, etc.

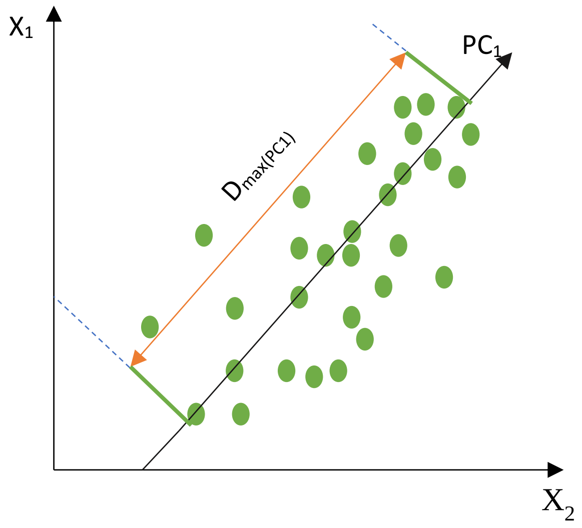


Figure 35 shows PCA space and the first principal component that defines the biggest variance in the dataset (image adapted from Wikipedia, by author Agor153, February 17, 2020, https://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%B9%D0%BB:FirstPrincipalComponent.jpg

Figure 35 shows a dimensional reduction of the data, where after applying the principal component method, the data presented in 2D space (axis X1 and X2) reduced its dimensionality to 1D space. The first principal component (PC_1) is orientated along the direction of the highest concentration of data points in a dataset. First principal component shows the greatest variance of the dataset. Unfortunately, we can see that data is scattered unevenly, and it means that with one principal component, it is impossible to define the dispersion of the whole dataset, hence other components need to be built. Altogether, these principal components will set up the dispersion for the entire dataset. It means that every principal component has its certain weight of the total variance of the dataset that is called loadings. The variance is a measure of data variability that may give valuable information about the content. Indeed, along with some PC axes, data variability can be high, along with some others, small or even absent. Principle components with higher variance will have a significant contribution in defining the model of a dataset. Principal components with a smaller variance will have less impact on the model; hence, they are assumed to be noise and can be discarded.

3.3. Cowe‘s example-based model for encoding facial dynamics

Glyn Cowe (Cowe, 2003) developed a technique for encoding facial dynamics and facial mimicry. His technique for mapping motion between faces requires two video sequences of faces in motion (i.e. driving and target sequences). The facial motion information of a person from the driving sequence will be extracted and mapped onto another person's face in the target sequence. According to Cowe, a bigger variety of non-rigid motions of a face in both sequences may result in more realistic reconstructions due to some overlapping of the facial mimics. Each image in both video sequences is presented in the vectorised format. The vector contains information about how the shape of a target face needs to be warped to cover a reference frame from the target sequence and the warped texture (i.e. the warped RGB image). Next, these vectors will be sorted with the principal component analysis (PCA) technique. This generates principal components that constituted a model of how the target face will vary. A detailed description of this technique presented below.

3.3.1. Image vectorisation

In the first step of Cowe's method, he expressed 2D images in 1D image vectors with an image vectorisation approach (Sirovich & Kirby, 1987; Turk & Pentland, 1991). With this approach, the image was represented by a h x w matrix X that held pixels' intensity (where h is height and w is the weight of image). Then, he placed rows of image matrix on top of one another and created a N x 1, 1D image vector x (where N is h x w) (see **Figure 36**).

Keeping in mind that Cowe's mapping method requires video sequences with M number of images, these video sequences will be vectorised as following $x_1, x_2, \dots\ x_M,$ then sorted and presented with PCA space. In addition, the initial vectorisation approach was applied only on images of one colour channel, but Cowe improved this approach and applied onto images with three colour channels.

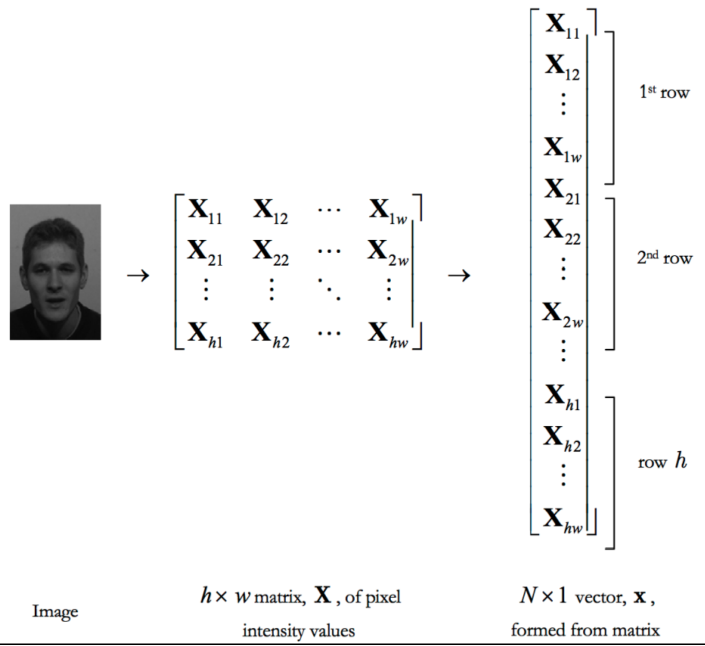

$\begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \cdots & \mathbf{X}_{1w} \\ \mathbf{X}_{21} & \mathbf{X}_{22} & \cdots & \mathbf{X}_{2w} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_{h1} & \mathbf{X}_{h2} & \cdots & \mathbf{X}_{hw} \end{bmatrix}$
1st row
2nd row
row h
Image
$h \times w$ matrix, $\mathbf{X}$, of pixel intensity values
$N \times 1$ vector, $\mathbf{x}$, formed from matrix

Figure 36 Vectorising 2D image. Image is defined by its pixel intensity values. N has size $\boldsymbol{h}$ **x** $\boldsymbol{w}$ (reproduced from Cowe 2003).

3.3.2. Centring images

In the second step, Cowe centred all frames of a sequence by initially chosen reference frame by using a mean Formulae. (4.0).

$$\boldsymbol{\mu} = \frac{\mathbf{1}}{\boldsymbol{M}} \sum_{i=1}^{M} \boldsymbol{x}_i \qquad (4.0)$$

He created a common space for all faces, and reference face was their mean $\boldsymbol{\mu}$. He had found a changing vector $\boldsymbol{\varphi}$ for every frame $\boldsymbol{x}$ in the sequence. Vector expresses how frames vary from a reference face image. After centring images, each frame $\boldsymbol{X}$ was expressed as a $\boldsymbol{\mu} + \boldsymbol{\varphi}$ (see **Figure 37**). Following Cowe's suggestions, reference face frame had to have common features that were present overall images.

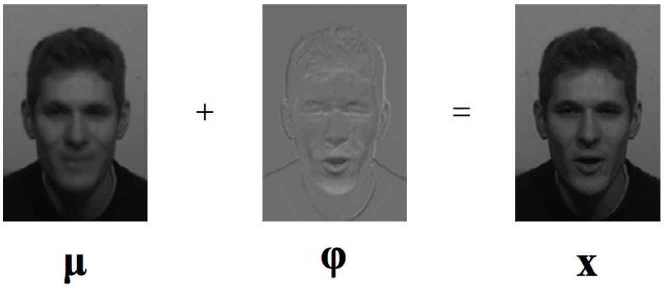


Figure 37 Initial face frame $\mathbf{x}$ is represented with sequence mean $\boldsymbol{\mu}$ and a change vector $\boldsymbol{\varphi}$ (reproduced from Cowe 2003).

3.3.3. Warping and morphing images

In the third step, images were warped and then morphed to achieve the most realistic face reconstructions, with a face mapping approach.

For the warping process, a reference frame $\boldsymbol{R}$ should be chosen, and all images in a sequence must be warped and defined based on this single frame. The warping process expresses a relation between pixels of a source image $\boldsymbol{R}$ and destination image, $\boldsymbol{T}$ by calculating a flow field, $[\boldsymbol{U}, \boldsymbol{V}]$, with the McGM algorithm (Johnston et al., 1999). Parameters $\boldsymbol{U}$ and $\boldsymbol{V}$ are components of a flow field that represents horizontal and vertical position of every (x, y) pixel. Components will be vectorised and centred, as previously explained in 3.3.1. and 3.3.2. Image wrapping was used to decrease the blurring effect of sharp edges in images that appear due to a linear combination of images. Hence aligning images by some constant face shape will help to minimise the blur.

Unfortunately, image reconstructions could not be fully accomplished with the warping process only because it fails to capture specific changes of the face. Thus, Cowe vectorised face images by morphing, where he combined warping and image blending processes. For the warping process, he mixed source and destination images to obtain a smooth transition between them. From the beginning, the weights of the contribution of the reference image had to decrease while coming closer to the destination image. The blending process looked like a source frame is fading on the background of the destination frame. Results of face reconstructions obtained with combined warping and image morphing are presented in (**Figure 38**).

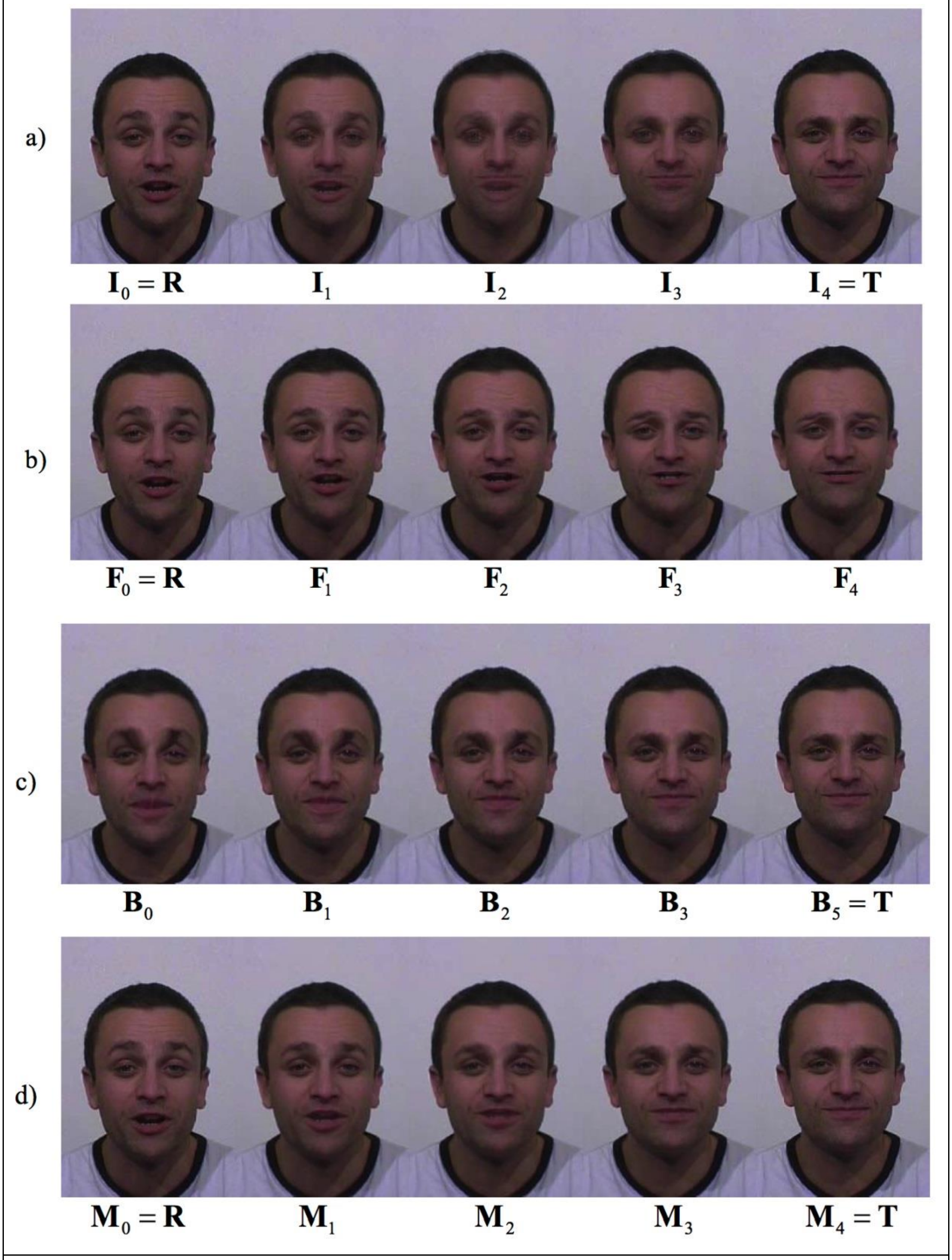


Figure 38 presents results of using different approaches for face reconstruction from one image into another a) blending approach that used to map reference image R into destination image T; b) warping approach from R to T; c) warping approach from T to R; d) results of combining of two approaches warping and morphing together to obtain smooth transitions between faces. (image reproduced from Cowe 2003).

After these steps, Cowe could represent images with shape and texture information concatenated to form one 1D long vector. Image shape information is obtained with a flow field $[\boldsymbol{U}, \boldsymbol{V}]$ that defines positions of features in the face, as $\boldsymbol{x}$ and $\boldsymbol{y}$ components in the vector. A texture is obtained from image morphing. When specific changes in the face could not be captured with warping, each frame and its face shape information will need to be reversely warped onto the mean face shape. It will allow to sort all the face features and leave information related to illuminance, occlusions or noise. Such data will be used as texture and expressed with R, G, B colour values for every pixel in the image (see **Figure 39**).

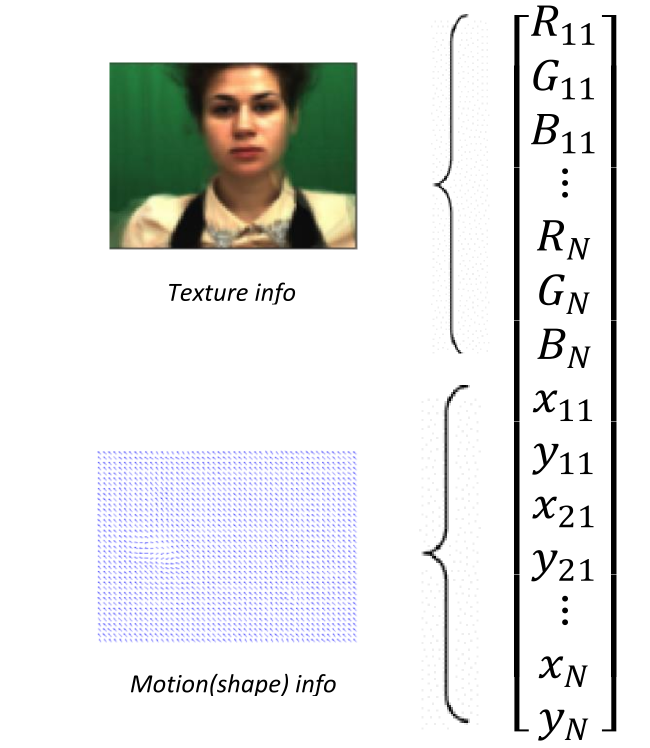

Texture info
Motion(shape) info
$\begin{bmatrix} R_{11} \\ G_{11} \\ B_{11} \\ \vdots \\ R_N \\ G_N \\ B_N \\ x_{11} \\ y_{11} \\ x_{21} \\ y_{21} \\ \vdots \\ x_N \\ y_N \end{bmatrix}$

Figure 39 One dimensional morph vector of the face. (image adapted from Cowe (2003)).

The above-listed steps of Cowe's method describe how to create a PCA face space for one individual with non-rigid facial movements. In the next part of his work, he tested whenever it is possible to move between face spaces by transferring non-rigid facial information from one individual onto another. Both individuals in video sequences were captured from the same viewing angle and said the same non-rigid information. His experiment in mapping between face spaces gave good results and was successful in transferring facial mimics between individuals (see **Figure 40**).

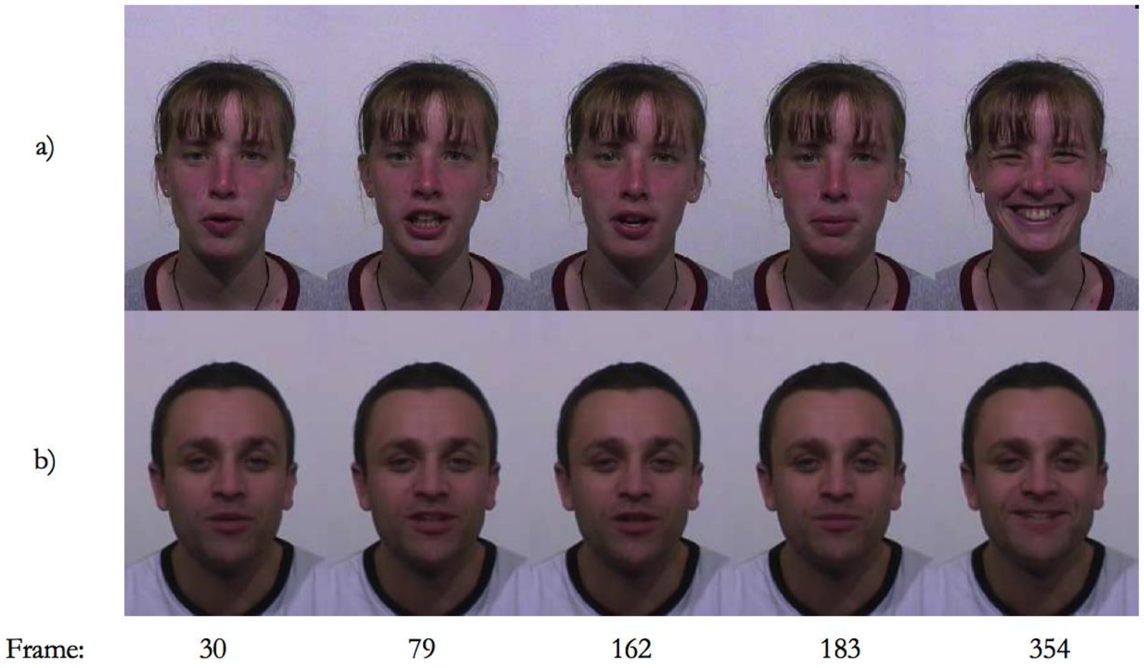


Figure 40 shows the results of transferring non-rigid facial information between individuals. a) frames from a source sequence. b) frames from a destination sequence that were warped with source frames and then projected into the space of individual in destination sequence (image reproduced from Cowe (2003)).

This method of mapping between spaces derived from different individuals will be generalised to map between models defined by pose, lighting or unfamiliar faces and will be shown in subsequent chapters.

3.4. Chapter 3 summary

Chapter 3 provided a detailed explanation of the hardware and software together with fundamental methods that were used to build the face system for this thesis. This system consists of appropriately configured hardware to capture facial movement and software that accurately calibrates data and performs PCA data analysis. The hardware consisted of a capture rig with six moveable cameras mounted on a rigid arc that produced high-quality images of non-rigid facial movements concurrently from multiple perspectives. These high-quality images were analysed with software that fixed image distortions arising from the cameras by finding the camera's relation with real-world units. Next, the alignment problem that arises across images captured from different perspectives was solved. Then, a model for automatically creating computer-generated avatars was integrated into the system. Such a model uses the face information of participants directly captured by cameras. In addition, we performed a face alignment procedure by aligning images by the nose bridge. Finally, aligned vectorised face sequences of a source were projected onto the destination face space. Results of creating computer-generated avatars with mapping non-rigid motions between two different faces were realistic and quite precise in facial motion reconstruction. However, the approach required to accomplish complex calculations consisting of several steps such as an image vectorisation, warping and blending. Keeping in mind a working process of this method, all the pros and cons, this model has been chosen and will be re-purposed to address the invariance over pose, lighting problem and unfamiliar face reconstructions. In the next chapters, empirical approaches to solving the problem of face recognition under various pose, illuminance conditions and face reconstructions for unfamiliar faces will be described.

4. Object constancy over multiple views from a PCA-based facial mimicry model

The way humans analyse object and face perceptions attracts a great deal of interest, mainly because of its many applications in a variety of fields such as psychology, security, computer technology, medicine and computer graphics. One of the major challenges that confront face reconstruction systems is how they should handle variations of the arbitrary poses. While different approaches have been developed for face discrimination across pose variations, many of these methods required manual landmark annotations (Beier & Neely, 1992; Kanade & Yamada, 2003; Lucey & Chen, 2008) or assumed the facial pose to be known (Blanz & Vetter, 2003; Fischer et al., 2012). Such constraints prevent many of the face perception systems from working automatically. In this chapter, a new approach is proposed, which is a partially automated method for obtaining face constancy over multiple views using only one example view. Face constancy over multiple views is an implicit shape recovery task. Although it is tied to the 3D shape of the face, it does not require an explicit computation of the 3D structure. The problem arises in the reconstruction tasks where a face, for which only one example image is available, needs to be reconstructed in a novel view. The few papers that have tried to solve object/face constancy problem investigated this problem within common face processing tasks like face recognition and tracking. As already described in Chapter 2, the majority of scientific research in face recognition obtained results by testing on facial data collected in controlled conditions, like with a small angle between viewpoints, or with a limited set of expressions (Beymer & Poggio, 1995; Blanz & Vetter, 2003; Fischer et al., 2012). Some of these research used PCA technique as a content addressable memory to recover information that was lost due to face occlusions (Jackson et al., 2017; Ding et al., 2016). This thesis will take it a step further and propose an alternative appearance-based approach for solving object constancy over multiple views, which builds on the existing PCA-based model developed by Cowe (Cowe 2003).

4.1. Multiple views representation on the Cowe PCA – based model

The primary goal of this thesis chapter is to use the existing system described in chapter 3 as a model of how people might recognise faces from different viewpoints, which may help to guide efforts to develop practical automatic face recognition systems.

In this empirical study, the system mentioned in chapter 3 uses PCA to generate example-based models of the face. The computational experiments are designed to explore a continuing research problem in computer vision – the problem of reconstructing faces from different perspectives. Specifically, all three experiments tested whether a Principal Components Analysis of the multiple views delivered simultaneously by six fixed cameras can by encoding global, correlated changes:

1) Distinguish rigid head movements of the head and non-rigid facial movements,
2) Reconstruct facial information seen from one perspective into another, and
3) Reproduce the facial information in multiple views from a single view.

The quality of the results of the animated experiment was examined by visually comparing it to the ground-truth representation in an attempt to reveal the quality of the reproduction. For experiment 3, the correlation coefficient was found to quantify the degree to which ground-truth and reconstructed results relate to each other. The reconstruction results of the PCA-based model will be compared with results of other work and presented at the end of this chapter.

4.1.1. Test stimuli: the database of multiple face sequences

A set of facial motion sequences were created to test a model. The mostly non-rigid facial behaviour was captured from different viewpoints at the same time and in the same lighting conditions as described in 3.1. Sequences were captured using six Grasshopper GRAS-03K2C (FireWire) digital cameras, at a rate of 60 frames per second, at 640x480 image resolution with an RGB 24 bit pixel format. The sequences were of an expressive talking face, and they lasted around 30 seconds. The capture conditions differed slightly in the different experiments. The experiment's sequences included:

- *1st Experiment:* An aligned, as described in 3.1.1., and calibrated, as described in 3.1.2., dynamic face sequence (non-rigid motion of the face from different perspectives) (see **Figure 41**, 1st row).
- *2nd Experiment:* An aligned and calibrated dynamic face sequence (non-rigid motion of facial behaviour from camera 1, camera 2 and camera 6) (see **Figure 41**, 2nd to 4th row).
- *3rd Experiment:* An aligned and calibrated dynamic face sequences (one serving as ground-truth, represents a multiple view capture of the face in non-rigid motion, while the other is missing all but one viewpoint in multiple views) (see **Figure 41**, 5th row).

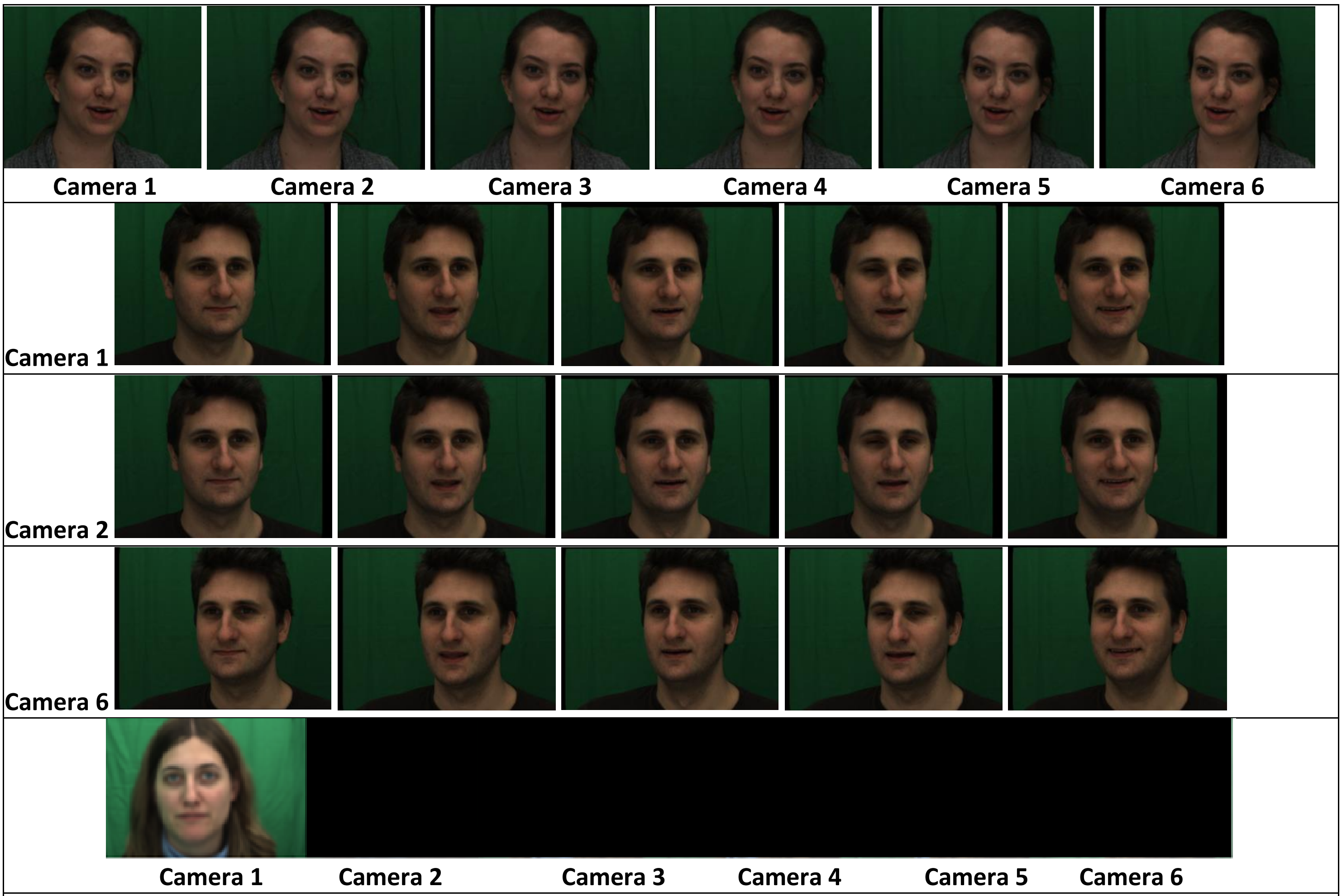


Figure 41 Examples of captured sequences from the database. **1st** row is from the 1st Experiment motion sequence (in total 260 frames per camera). **2 nd** to **4th** row is from the 2nd Experiment motion sequence. The **5th** row is from the 3rd Experiment's motion sequence (concatenated multiple views with all but one viewpoint missing).

4.1.2. The Experimental results

The sequences were vectorised as described in 3.3.1. The sequences for each face were processed using PCA, and the basis vectors were extracted, forming a model of the target face based on principal components.

4.1.3. Experiment 1

The results of the first experiment reveal the principal components that define the rigid motion of the head and non-rigid movements in the face region over the multiple-view sequence.

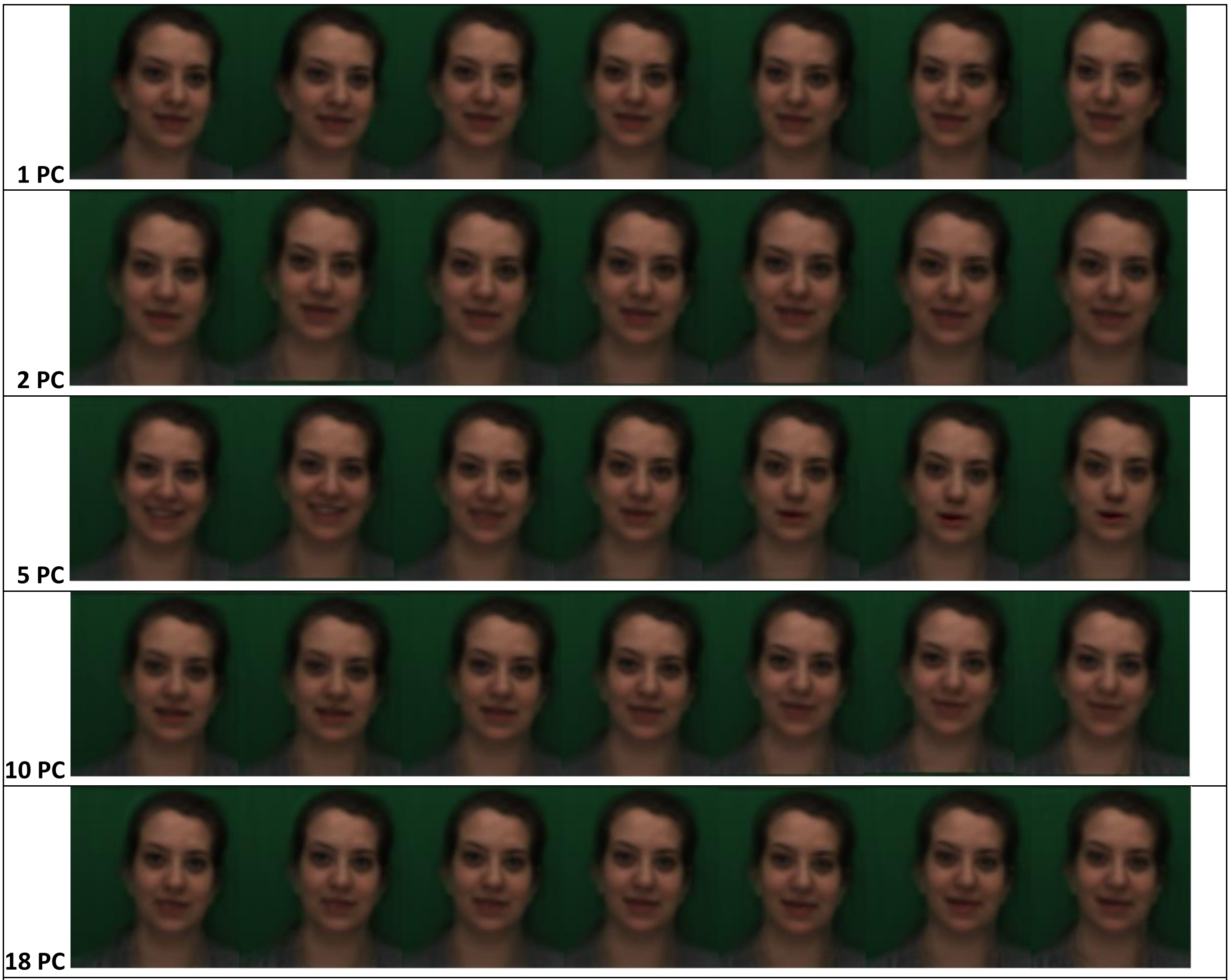


Figure 42 The figures show reconstructed images for various values of the loadings on the components arrayed in a row. The first component ('1PC') reflects the rigid movement of the head. It contains most of the variance in the sequence because it captures the pose of the head. The second component reflects a rigid motion of the head, in which a squeezing movement can be seen over the face region, while the face expression stays the same. 5PC, 10PC and 18PC reflect non-rigid facial movements over the face, resulting from a change in expression (see **Movie 1.avi**).

From the multiple-view sequence, 18 principal components were obtained that characterised both rigid movement of the head and non-rigid movements over the face region. Images in the rows represent reconstructions of the face for various coordinate values of the principal component. From visual inspection, it would appear that the first four principal components correspond to the rigid motion of the head, while the rest correspond to non-rigid facial movements. The sequence of frames shown in **Figure 42** depicts the principal components that show the most information. Since the 1st PC contains most of the variance, and the variance decreases as we approach the 18th PC, the standard deviations of the principal components in **Figure 42** were increased to allow image variations to be seen clearly. The row labelled '1PC' contains images with most of the variance, which corresponds to the pose, and this is why we see an apparent rotation of the head. The second component contains information regarding head compression, while facial expression stays unchanged. The fifth component best describes the non-rigid movements of the face. In **Figure 42**, the row labelled '5PC' shows that the movement of the whole face is captured. It is unlike the other principal components that define the non-rigid facial movement, each of which contains only information regarding the movement of the specific parts of the face.

4.1.4. Experiment 2

The results of the second experiment show the mapping of the face from one perspective to another. See, for instance, how in **Figure 43** and **Figure 44**, a PCA space was computed for each of three camera views separately, specifically for camera 1, camera 2 and camera 6. The reference frame was taken from the same time point for all three sequences. To reconstruct the face from one viewpoint in another viewpoint, a difference vector was taken (frame was subtracted from the mean) in the face-space of the first camera viewpoint and projected it into the face-space of the second camera viewpoint and then reconstructed the image from the PCA space of the second camera (**Figure 43**). The frames shown in **Figure 43** are the results of this face reconstruction process for three principal components. From visual inspection, it can be seen that the reconstructed face is very accurate and close to the avatar frame obtained from the 2nd camera viewpoint, i.e. close to the ground-truth. This degree of accuracy might be expected because the 1st camera viewpoint differs by only 5 degrees from the 2nd camera viewpoint.

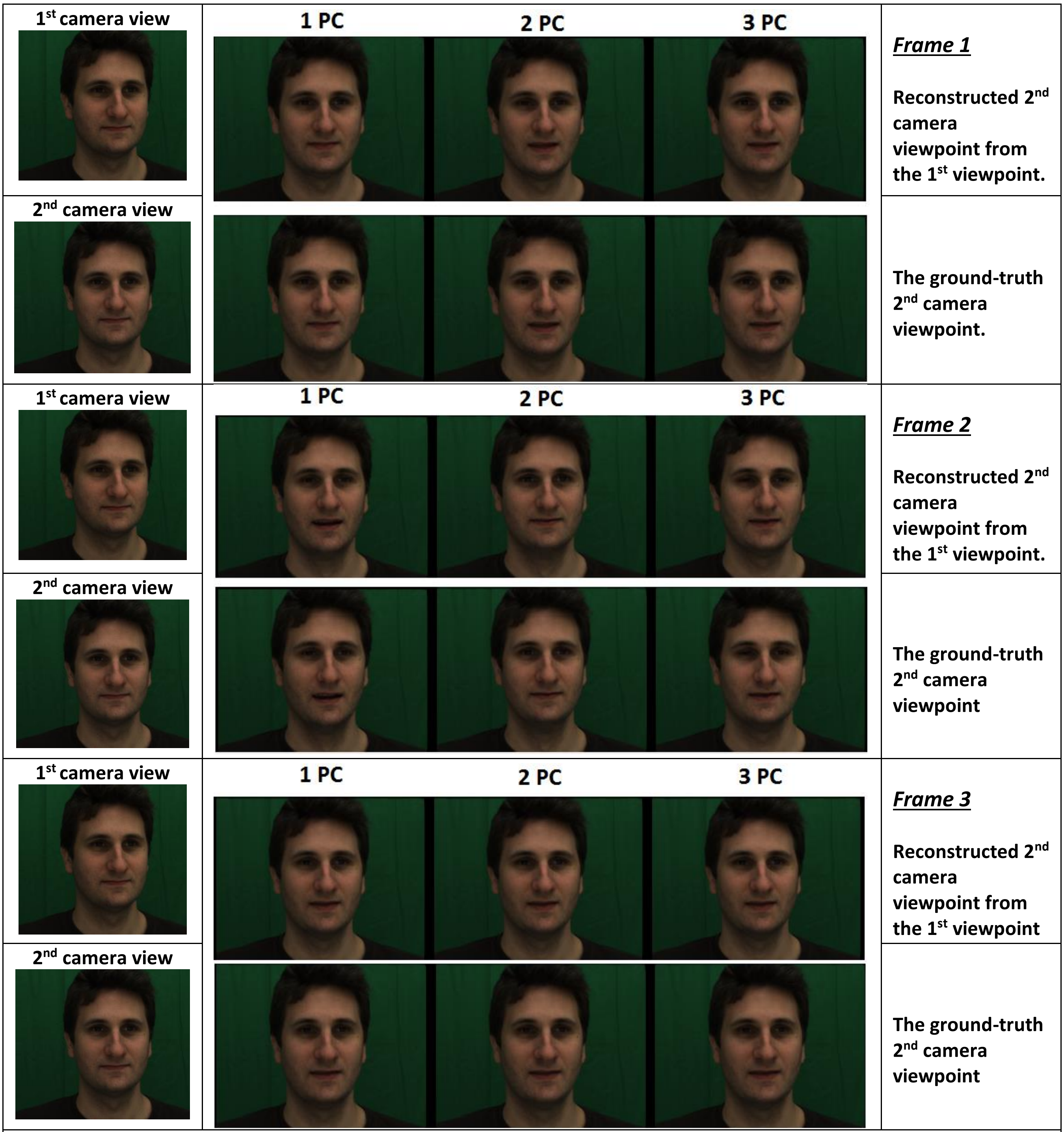

1st camera view
1 PC
2 PC
3 PC
Frame 1
Reconstructed 2nd camera viewpoint from the 1st viewpoint.
2nd camera view
The ground-truth 2nd camera viewpoint.
1st camera view
1 PC
2 PC
3 PC
Frame 2
Reconstructed 2nd camera viewpoint from the 1st viewpoint.
2nd camera view
The ground-truth 2nd camera viewpoint
1st camera view
1 PC
2 PC
3 PC
Frame 3
Reconstructed 2nd camera viewpoint from the 1st viewpoint
2nd camera view
The ground-truth 2nd camera viewpoint

Figure 43 *Frame 1*, *Frame 2* and *Frame 3* show the reconstruction of different facial expressions in the 2nd camera viewpoint from face information obtained from the 1st camera viewpoint. 1PC, 2PC and 3PC are the first three independent principal components from a sequence of facial motion vectorised as morphs.

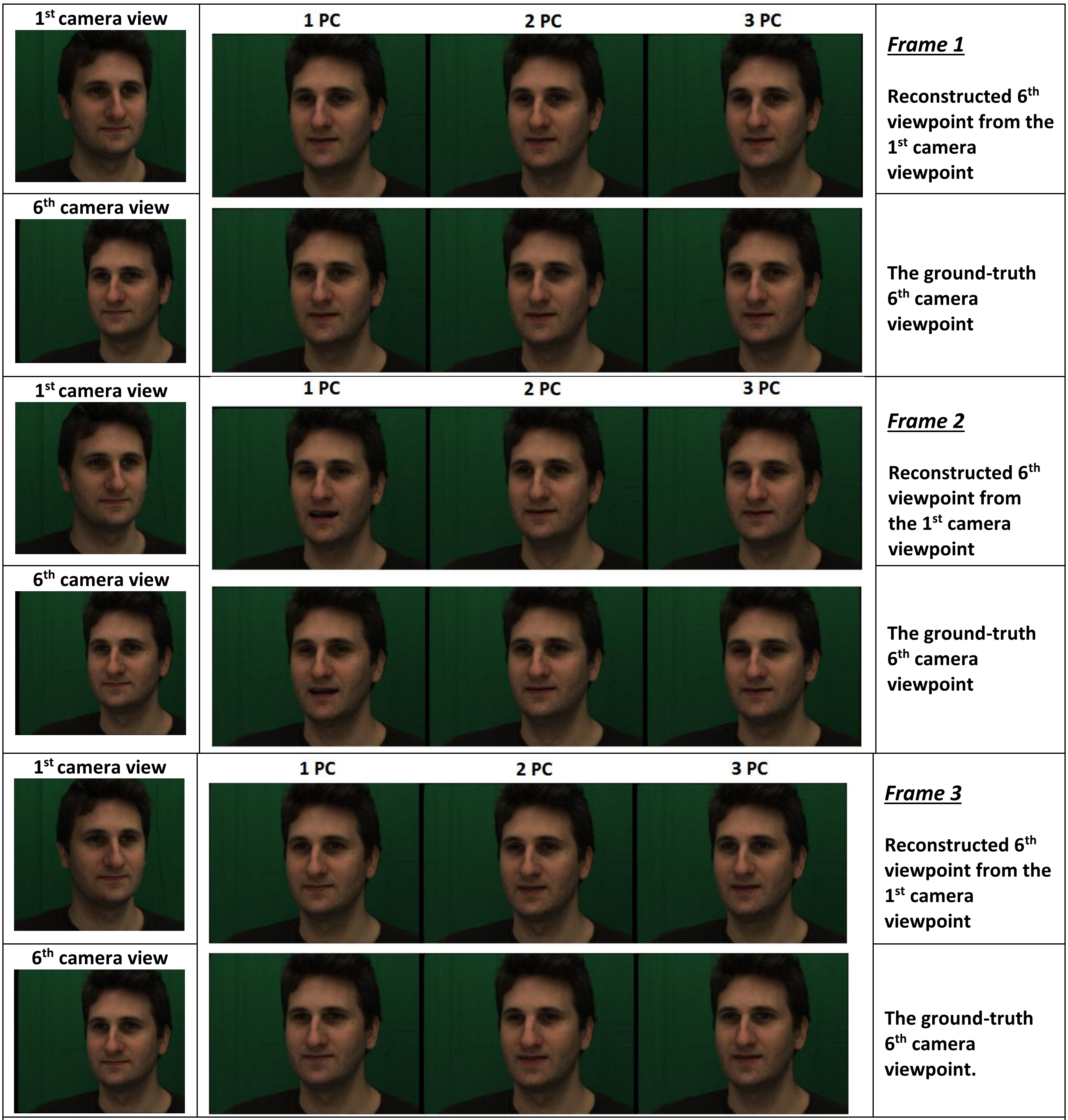

1st camera view
1 PC
2 PC
3 PC
Frame 1
Reconstructed 6th viewpoint from the 1st camera viewpoint
6th camera view
The ground-truth 6th camera viewpoint
1st camera view
1 PC
2 PC
3 PC
Frame 2
Reconstructed 6th viewpoint from the 1st camera viewpoint
6th camera view
The ground-truth 6th camera viewpoint
1st camera view
1 PC
2 PC
3 PC
Frame 3
Reconstructed 6th viewpoint from the 1st camera viewpoint
6th camera view
The ground-truth 6th camera viewpoint.

Figure 44 _Frame 1_, _Frame 2_ and _Frame 3_ show the reconstruction of different facial expressions in the 6th camera viewpoint from face information obtained from the 1st camera viewpoint. 1PC, 2PC and 3PC are the first three independent principal components from a sequence of facial motion vectorised as morphs.

For a more stringent test, the process was repeated with a much larger angular difference in the viewpoints (25 degrees). This mapping was performed as explained above, between the 1st and 6th camera viewpoints and is shown in **Figure 44**. From visual inspection, the facial reconstructions shown in **Figure 44** appear identical to the ground-truth. However, if we take a closer look at 'Frame 2', we can see that the mouth of the subject is opened slightly wider than on the ground-truth frame for that expression. Nevertheless, except for this small reconstruction error, reconstruction is detailed and precise.

4.1.5. Experiment 3

Berisha, Johnston, & McOwa (2010) reconstructed a full face from an occluded version of the face using the Cowe technique (Cowe, 2003). In this case, a partial vector constructed from the test face is projected into a PCA space derived from a facial action sequence for that face. They showed that the facial expression could be recovered from unoccluded regions and that the mouth and eyebrows regions contained most information about the global structure of the face. They only used frontal views of the face.

Can this methodology be used to understand how people might build multiple view-based representations that could be used to support generalisation over pose?

High quality, high frame rate videos were recorded of the moving face from six cameras simultaneously, with relatively large camera separations (18 degrees) in a horizontal arc around the face. The synchronised videos sequences from each perspective were converted into separate .png image files and concatenated to form a panoramic multi-view representation of the face for a particular point in time. In this experiment, only one PCA face-space containing all the information presented was created. PCA will be used as a content addressable memory (Cowe, 2003; Kohonen, 1986). Projected part of the content addressable memory can recover information stored in the PCA space. This property will be used to recover all views from a limited set of views.

First, a multi-view vector is constructed for each time step. By multi-view, it means that multiple view images with the same expression were concatenated into one image, and, after that, these multi-views were treated as a single unit. The content addressable memory is loaded with 446 multi-views to include a wide range of non-rigid facial expressions. Then this PCA space is used to recover the ground-truth (*multi-view* **Figure 45**) information by projecting the full vector into space and the missing viewpoints by projecting a partial vector containing just some of the views (*multi-view* **Figure 46**). The first step of face reconstruction was removing the information about all but one view from the multi-view vector and setting the vector information derived from the other views to zero **(Figure 46)**.

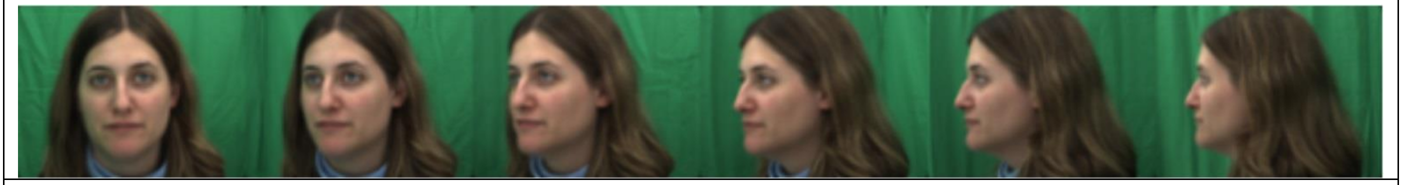


Figure 45 presents one of 466 concatenated multiple views, which are stored in the content addressable memory.

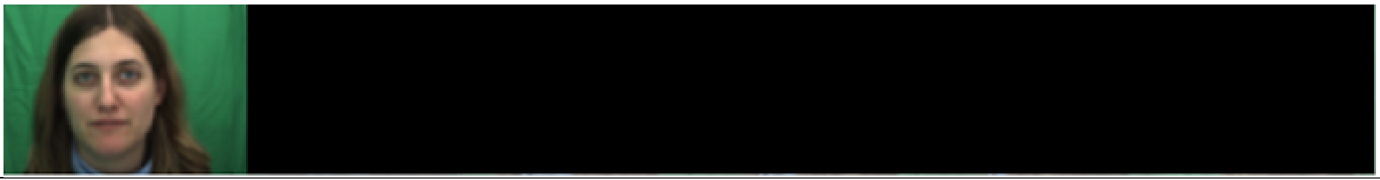
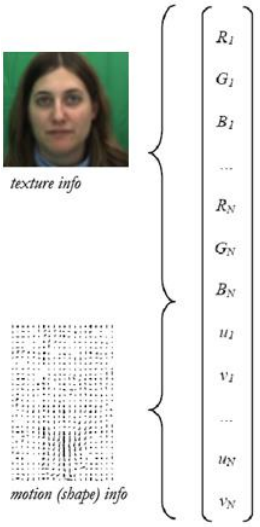


Figure 46 a concatenated multiple view vector schematically with the information from all but one viewpoint removed.

The second step was projecting this partial representation back into the principal component model to regenerate the other views **(Figure 47)**.



$$\begin{array}{cccccc} \text{Recon} & & & \text{PC1} & & & \text{PC2} & \\ \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ \vdots \\ r_n \end{bmatrix} & = & \begin{bmatrix} t_1 & t_2 & 0 & 0 & 0 & \cdots & t_n \end{bmatrix} & \begin{bmatrix} c1_1 \\ c1_2 \\ c1_3 \\ c1_4 \\ c1_5 \\ \vdots \\ c1_n \end{bmatrix} c1 + & \begin{bmatrix} t_1 & t_2 & 0 & 0 & 0 & \cdots & t_n \end{bmatrix} & \begin{bmatrix} c2_1 \\ c2_2 \\ c2_3 \\ c2_4 \\ c2_5 \\ \vdots \\ c2_n \end{bmatrix} c2 + \cdots \end{array}$$

Figure 47 presents a schematic view of the formula, which is required to reconstruct the ground-truth image from only one perspective. A partial vector in which the zero elements represent the views we wish to recover is projected onto the principal components, both of which have the structure as shown on the left of the figure. The inner product of the partial vector and the principal component is the scalar weight, which determines the proportion of that component to be added into the reconstructed vector. The reconstructed image is then computed from the reconstructed vector.

The views of the face reconstructed from only one view reproduced the pose of the subject's head; however, it was not able to reproduce the facial expression in the ground-truth as seen in **Figure 48**.

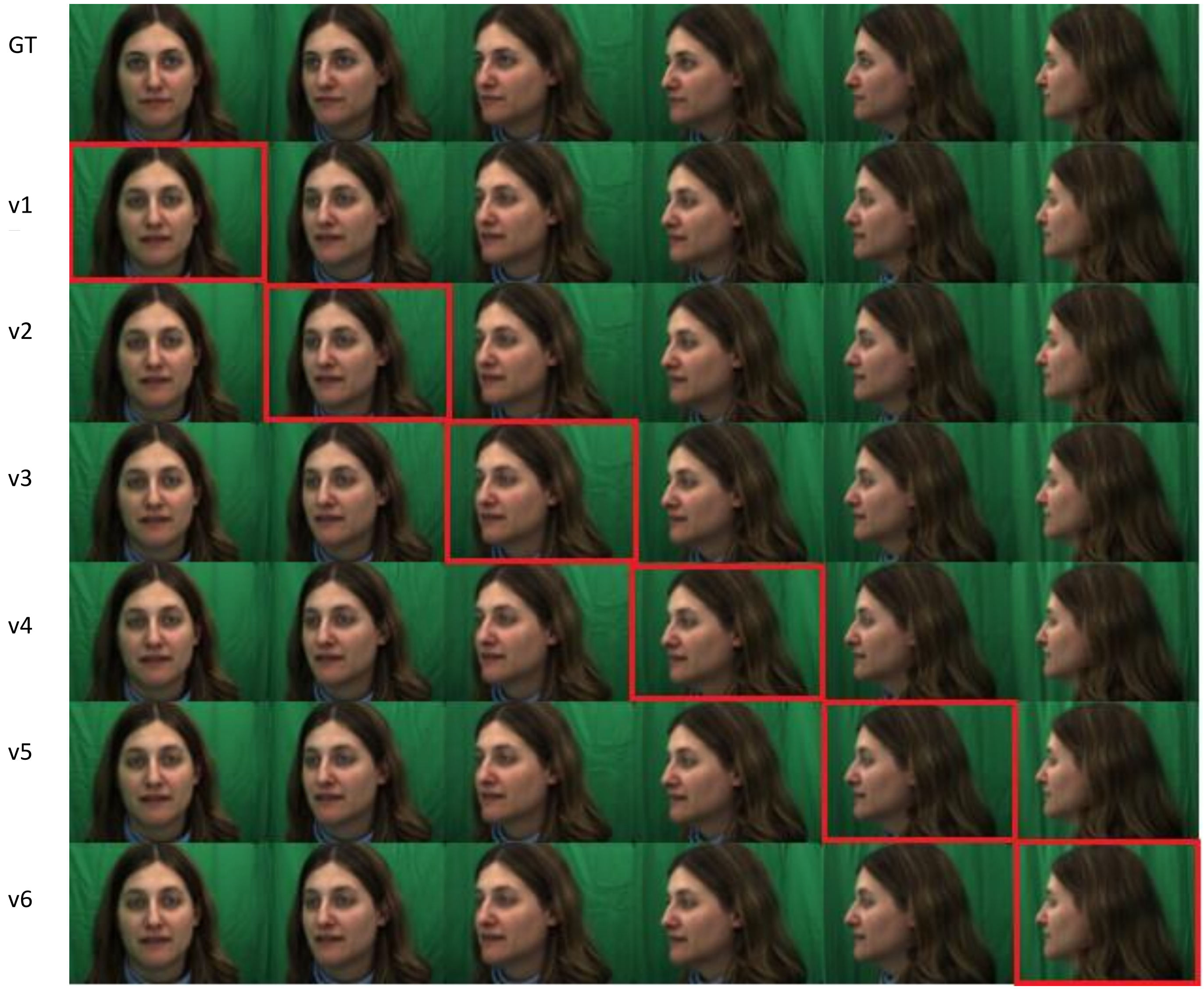


Figure 48 presents the ground-truth image and the reconstructed views using only information from one perspective. The red bounding box depicts which view of the image was used to reconstruct the face in other views. The reconstruction in all views (v1,v2,v3,v4, v5,v6) looks very different from GT image representation (see **Movie 2**.avi).

With the naked eye, it is tough to determine precisely how different the reconstructed views are from the ground-truth representation. Hence, we have plotted the weights multiplying the 1st principal component for both the reconstructed and ground-truth data in the face space **(Figure 49).**

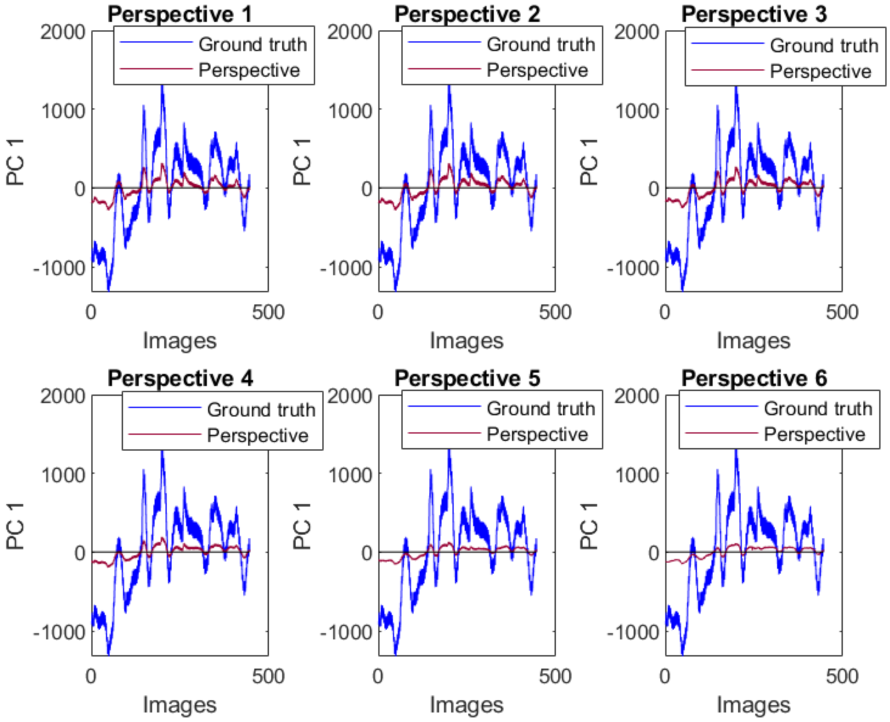


Figure 49 depicts 1st PC weights for the ground-truth image (blue graph) and weights for the reconstructions from the 1st, 2nd, 3rd, 4th, 5th, 6th views (red graph). The “reconstructed” (red graph) values are very small (notice the images in **Figure 48**). Hence a gain factor will be calculated to make both graphs more similar.

It is clear from the representation of the 1st principal component that the values of the reconstructed image from all views (red graph) are much smaller than ground-truth values (blue graph). Therefore, we have chosen to enhance the reconstructed values to match the ground-truth representation. By enhancing reconstructed values, we will have a better view to observe consistent facial movements and delicate non-rigid changes of a face that were barely noticeable in the reconstructed sequence. A linear least-squares method was used to find a scale factor that brings the two curves into correspondence. To omit data overfitting, only 20 per cent of all trials of the database was used to find a scale factor. For the current case, when the database consists of 446 data inputs, only the first 90 data inputs were taken to obtain a scaling factor. The scaling factor was calculated for every principal component (ten principal components) and views, and the results are shown in **Figure 50.**

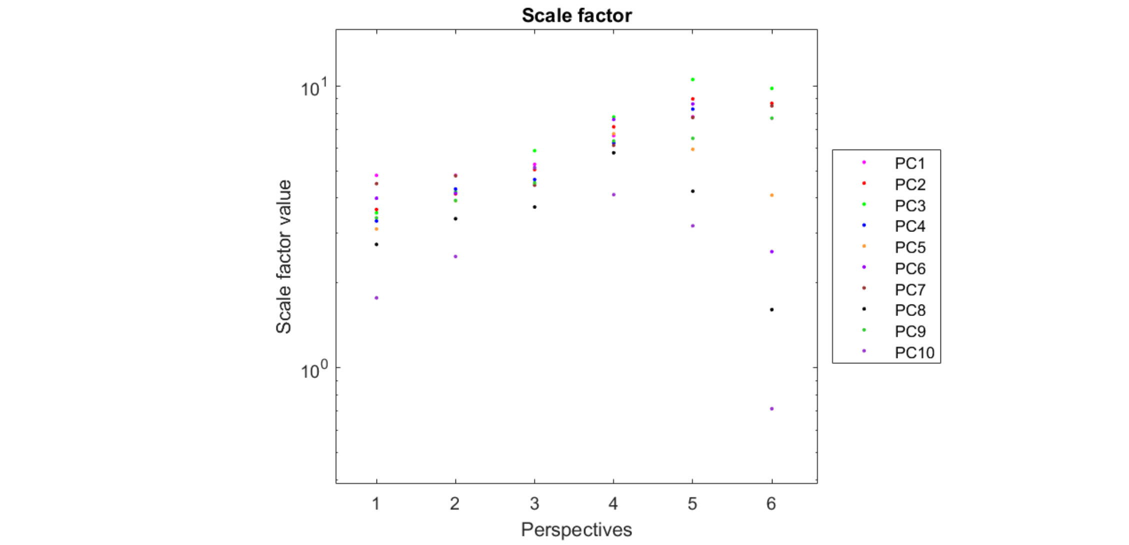


Figure 50 shows the scale factor values obtained via a least square method for finding the best match between the GT and reconstructed images principal component values for all frames in the sequence.

Figure 50, the graph shows the data for the reconstructions derived from the labelled view. Here we may see that the range of the scale factor derived for the 3rd view is much narrower than the scaling factor derived for the 6th view. That is presumably due to the 6th perspective being a profile face, holding less dynamic facial information that is present in the 3/4 facial representation captured from the 3rd viewpoint. Obtained scaling coefficient from the reserved part of dataset was applied onto the whole dataset, that is composed of 446 reconstructed values, to test how reconstructed values match the ground-truth.

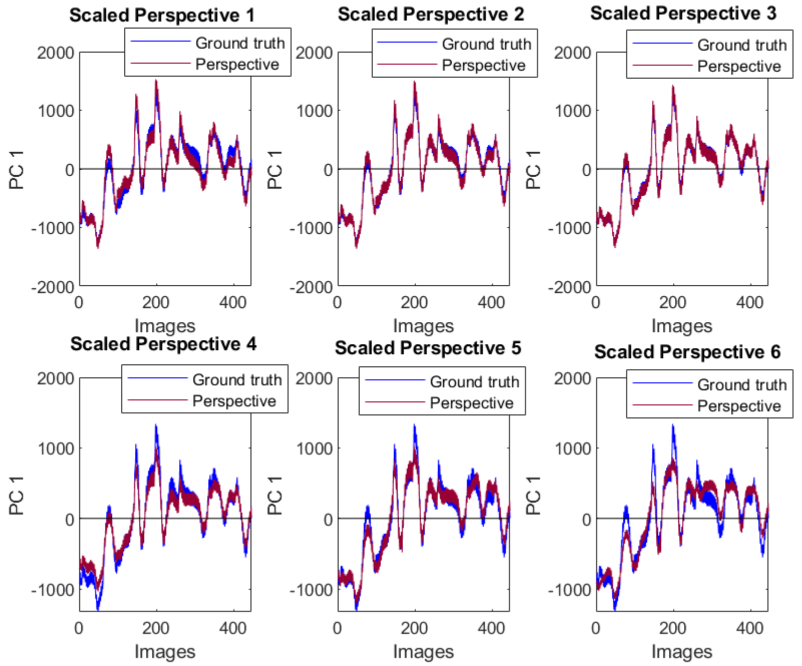


Figure 51 represents the 1st PC weights for the ground-truth image (blue graph) and the values of the reconstruction from 1st, 2nd, 3rd, 4th, 5th, 6th views (red graph) after applying the computed scaling factor. The “reconstructed” (red graph) values became very close to the GT curves, which indicates that scaled reconstructed images should be very similar to GT image representation (blue graph).

Figure 51 shows the 1st principal component weights for the ground-truth image (blue graph) and the equivalent values for the reconstructed images from the 1st, 2nd, 3rd, 4th, 5th and 6th views (red graph) after applying a scaling factor, respectively. The scaled principal component weights for the reconstructed images resemble the ground-truth representation. The reconstructed images after applying the scaling factor are shown in **Figure 52**.

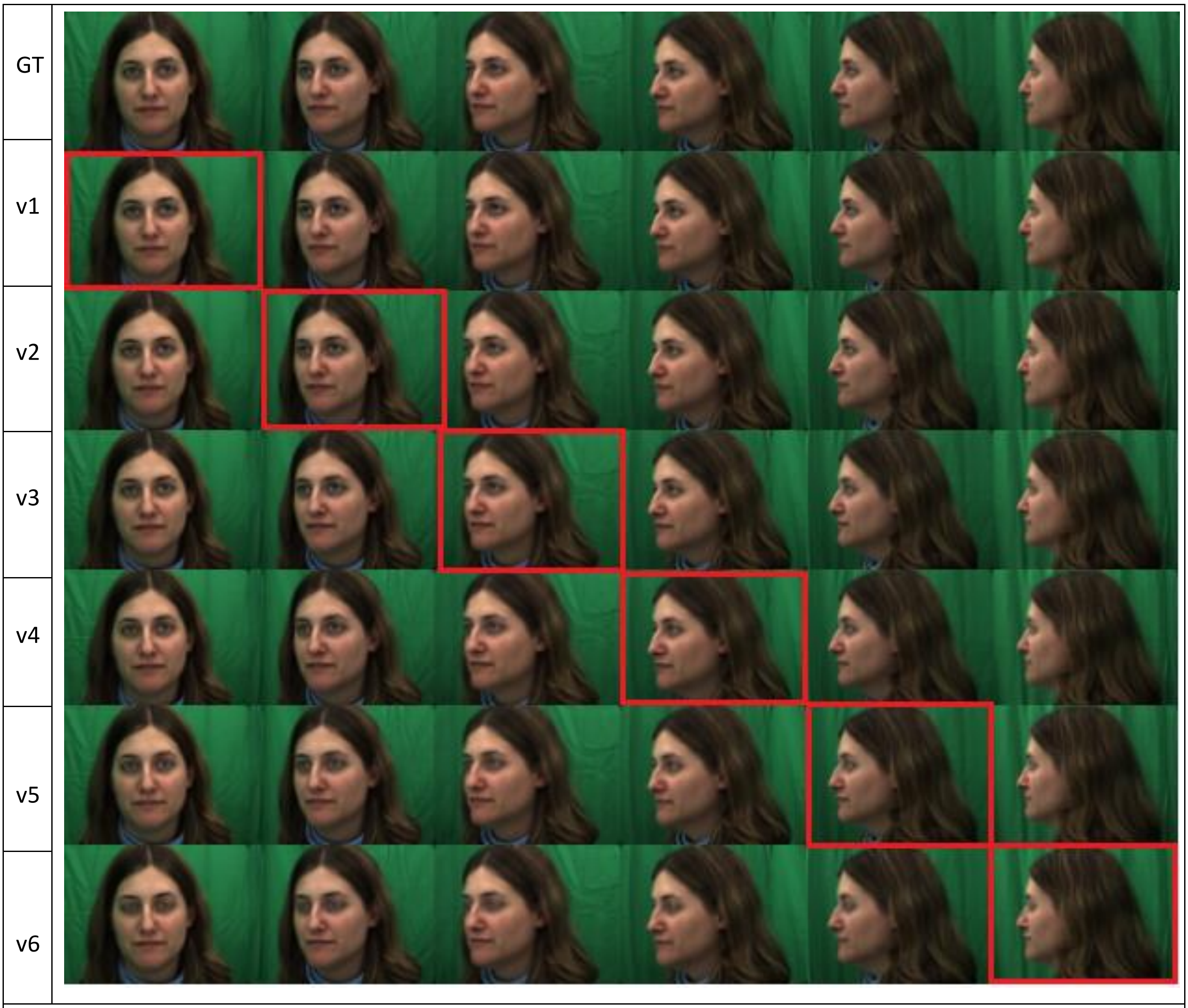


Figure 52 depicts a ground-truth multi-view image and reconstructed multi-views after applying a scaling factor. Red bounding box depicts view of the image loadings was used to reconstruct the face in other views. From visual inspection, the facial reconstruction in all views (v1,v2,v3,v4,v5,v6) appear identical to GT image representation (Movie 3.avi).

From visual inspection, it can be seen that the facial reconstruction for all views after applying a scaling factor looks very accurate and close to the ground-truth representation stored in the content addressable memory. The accuracy was to be expected since the scaling factors have been obtained explicitly for each principal component and view independently. To measure how well reconstructed and scaled multi-view images match the ground-truth image mimicry, a standard Matlab Pearson product-moment correlation coefficient approach was employed. It will be used to measure the value of similarity between the two multi-views. The values will be presented in the range from +1 and -1 inclusively, where 1 indicates that two multi-views are the same, 0 is no similarity, and -1 indicates a perfect negative similarity. At first, we used this approach in comparing RGB colour values of every pixel in every image, between scaled reconstructed and ground-truth, multi-view facial expressions. However, such comparison could not pick up on small image changes (place around the mouth), and high correlations did not correspond well to similarity as judged by inspection. A consequence of this, Pearson correlation was tested between loadings of principal components extracted from the 446 reconstructed multi-views for every of six views sequence vectors and those from the ground-truth. Correlation results corresponded well with results from visual inspection. Thus, correlation results are trustworthy and can be used in evaluating reconstruction performance. The results of measuring the dependence between scaled multi-views loadings and ground-truth are represented in **Figure 53**.

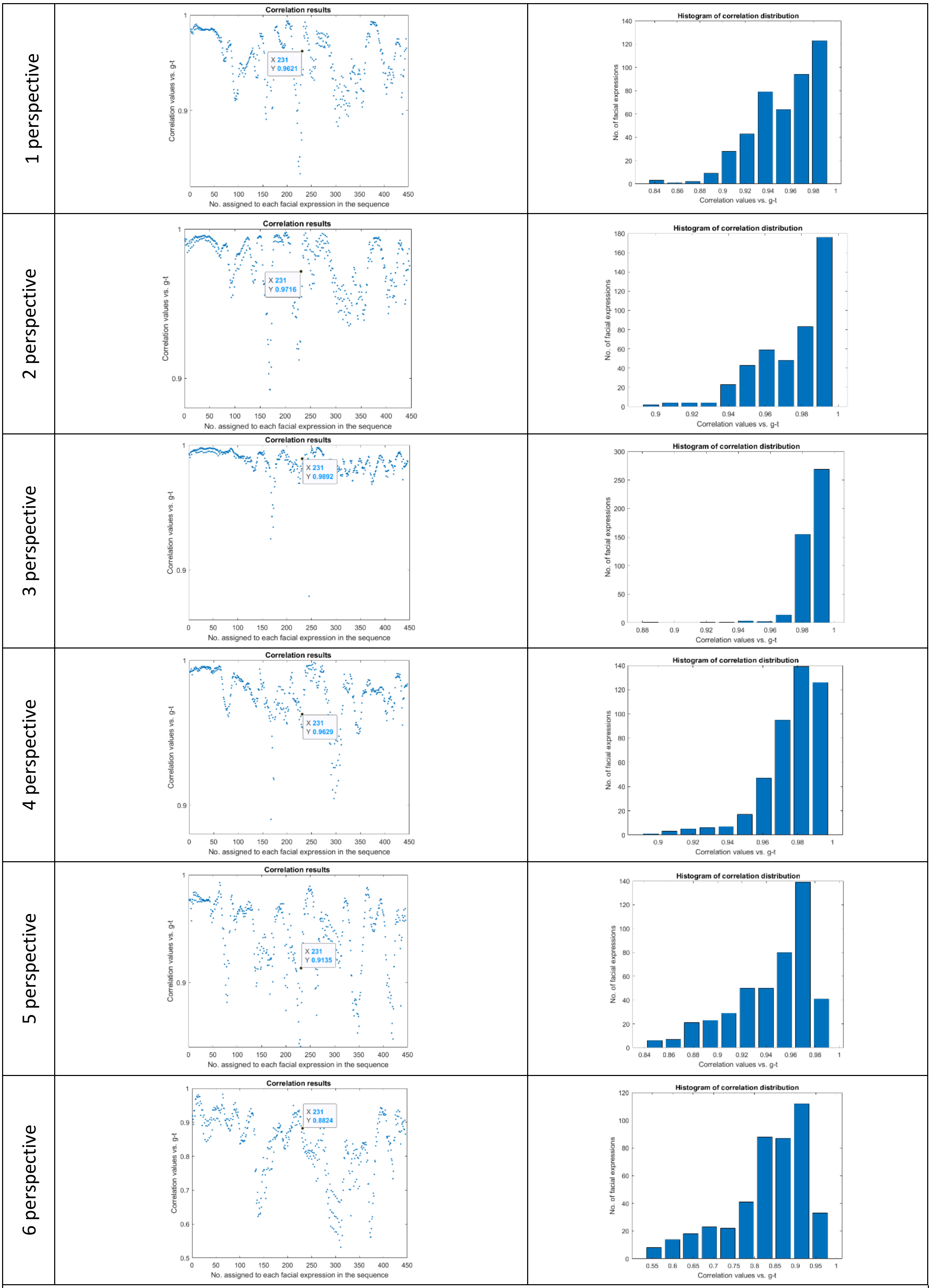


Figure 53 Correlation results between reconstructed multi-view images and ground-truth representation. In the graphs of the left column, we analysed loadings sequentially and separately for every view with loadings of ground-truth. For example, we have chosen a random multi-view face expression at position 231 and compared the correlation coefficients of each reconstructed view. We can obtain the highest reconstruction from only one view value, 98.92%, in case of using 3rd face posture information (Reconstructed from 3rd view; the third row from the top) and less information using profile face posture (Reconstructed from 6th view; bottom row). The histogram in graphs of the right column show correlation distribution for each reconstructed perspective.

The graphs of the left column in **Figure 53** show the correlation between the loadings of the reconstructed and scaled images with loadings of the ground-truth. A correlation value was obtained for every view of multi-view by taking loadings of multi-view from a single view of the reconstructed image and loadings of multi-view of the ground-truth, respectively. This process was repeated for every viewpoint. Also, we have randomly chosen a multi-view face expression from the sequence positioned at 231 to compare correlation coeffects of reconstructed multi-views. It can be noticed that the overall perspectives of the reconstructed images gave a very high correlation value of over 0.98. The highest face reconstruction with a 98.92 per cent value has been obtained using 3rd face view information. The smallest face reconstruction value was 88.24, and it was obtained using profile face information. Histogram graphs of the right column represent how correlation data is distributed for all reconstructed six multi-views. We may observe that correlation values for the first five reconstructed multi-views are all above 0.8 when correlation values for the profile view started from 0.5. However, we checked how many data from the profile view had a correlation value above 0.8. Calculations showed that 72,42 per cent of reconstructed data from the profile view had a correlation value above 0.8.

4.1.6. Method comparison

To have a better understanding of how good this method's results are, let us compare the PCA method with a well-known image morphing method based on "Feature-Based Image Metamorphosis" by Thaddeus Beier and Shawn Neely (Beier & Neely, 1992).

For fair results, both methods (i.e. morphing method and methods in 4.1.4.) will be tested on the same dataset. The testing performance of the morphing method is shown in **Figure 54**. In Frame 1 and Frame 2, from 1st camera view into 2nd camera view, destination (interpolated image) face image is reconstructed with good precision, including a non-rigid motion of the face. However, we may observe some ghosting effect outside the face area (ears, hair, and neck). It appears to be due to the different shape of the faces. A face has a different shape in various viewpoints. In this case distance between camera views reached only 5 degrees.

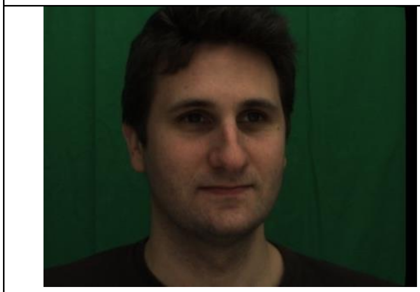
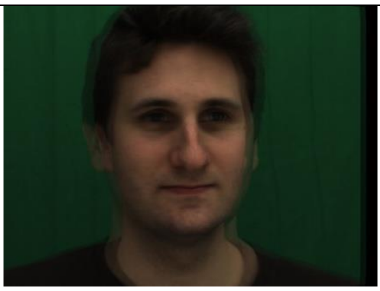
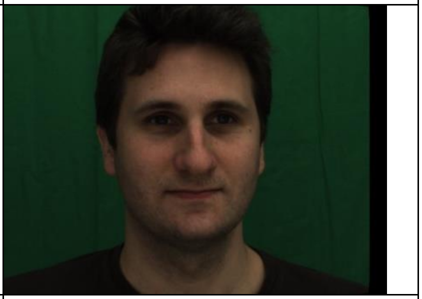
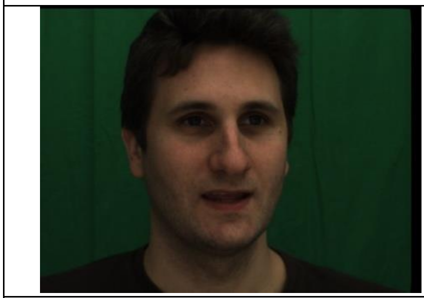
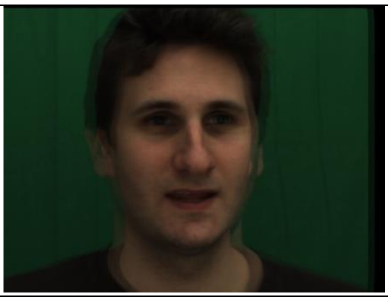
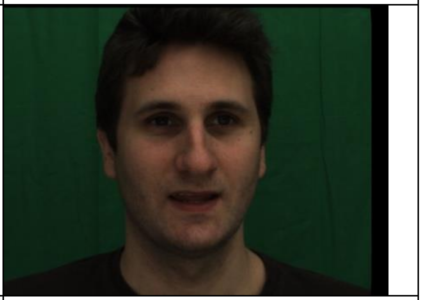
Frame 1		
1st camera view	**Interpolated image**	**2nd camera view**
		
Source image	**Destination image**	**Source image**
Frame 2		
1st camera view	**Interpolated image**	**2nd camera view**
		
Source image	**Destination image**	**Source image**

Figure 54 Frame 1 and Frame 2 shows the reconstruction results of different facial expressions using the morphing method, from the 1st camera viewpoint into the 2nd camera viewpoint. Destination image shows artefacts around the face area. It is due to the difference in the shape of the face.

Next, the morphing method has been tested with an enlarged angle between the cameras up to 25 degrees (see **Figure 55**).

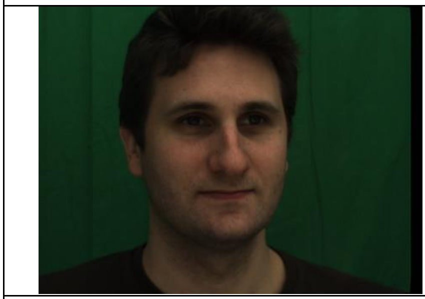
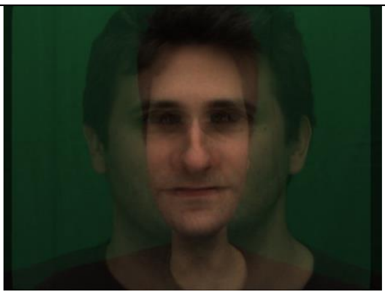
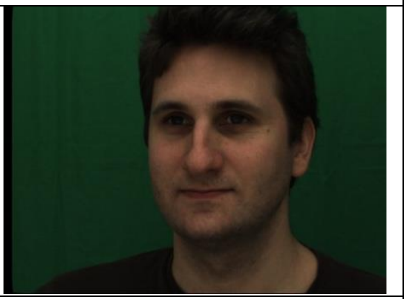
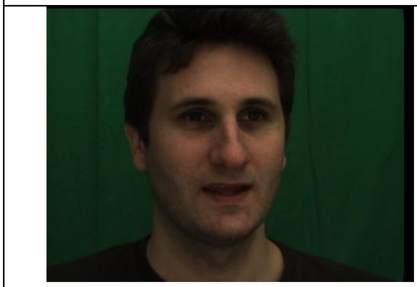
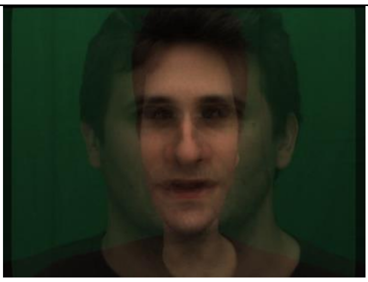
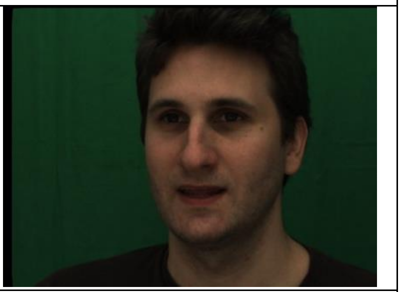
Frame 1		
1st camera view	**Interpolated image**	**6th camera view**
		
Source image	**Destination image**	**Source image**
Frame 2		
1st camera view	**Interpolated image**	**6th camera view**
		
Source image	**Destination image**	**Source image**

Figure 55 Frame 1 and Frame 2 shows the reconstruction results of different facial expressions using the morphing method, from the 1st camera viewpoint into the 2nd camera viewpoint. Destination image shows artefacts around the face area. It is due to the difference in the shape of the face.

The reconstruction results represented in the destination image appeared to be very poor. The ghosting effect around the face area became more significant and surpassed the face area. On the other hand multi-view method in 4.1.5 can map faces that that has a 90-degree angle between poses. As it was mentioned before, a wider angle between cameras results in a poorer destination image. Hence, a morphing technique will not give us a sharp and clear reconstruction of the face. With a wider angle between cameras views, the ghosting effect around the face increases.

To conclude methods comparison, we may say that it is evident that the warping method will work as Beymer and Poggio has shown in their study. However, it performs best within a few degrees of rotation, and it will fail badly with a vast difference in face pose (Beymer & Poggio, 1995). Unfortunately, with the morphing method, it is impossible to warp from the full-face view to the profile view.

4.2. Chapter 4 summary

The three experiments described in this chapter demonstrated that the PCA approach for generating and driving photo-realistic faces is a powerful tool for reconstructing and reproducing non-rigid facial information over multiple views. When pose and expression are incorporated in the same PCA space, the principal components can distinguish between rigid and non-rigid movements of the face, with the pose of the face accounting for the highest variance in this case (see 4.1.3.). The second experiment showed how the human visual system might deal with recognising faces in different poses (see 4.1.4.). It appears that we do not need to learn and store multiple face images from all different perspectives in our head to perform accurate recognition of the face as we can map between separate representations for different views of the face. In the third experiment, it has been shown that it is possible to reconstruct missing viewpoints from a multi-view representation with high precision (see 4.1.5.). This method of recognising/reconstructing faces from different perspectives with all but one viewpoint missing makes use of a simple statistical model that could potentially be used in the human visual system. These three experiments demonstrated the feasibility of object representation in a human visual system based on a translation between view-based models rather than a full 3D representation (Marr & Nishihara, 1978) or simple view association (Miyashita, 1993). To see how well the PCA method performs, it has been compared with the results from the morphing method. Both techniques (i.e. morphing method and method in experiment 2) had proven to bring good results when the angle between views was small (5 degrees). However, morphing results failed severely when the angle between views increased up to 25 degrees, and only the PCA approach was able to obtain high reconstruction results with minor defects. A multi-view face reconstruction approach can map frontal view to profile view, which makes 90 degrees angle between poses (experiment 3). Keeping in mind that morph method performance was abysmal when the angle between face poses reached 25 degrees, we can state that morph results will not handle face reconstruction in such large poses. Therefore the PCA method in both experiments performed better than morphing technique.

5. Object constancy under various illuminations facial mimicry model.

In the previous chapter, the problem of face constancy in multiple poses was introduced. This problem is caused by the fact that changes in the orientation of the head generate huge differences in the image. Another well-known variable in face perception that has a strong influence on facial images is the direction of the lighting. Images of the same person appear to be dramatically different under various light conditions. The direction of the light has a more significant effect on the facial image than changes in personal identity. Braje et al. confirmed such a statement in their research study where participants had to match a laser-scanned unfamiliar face with direct light applied only on one side of the face, with an unfamiliar face with direct light applied only on the other side of the face and decided if these scans represented the same face (Braje et al., 1998) (see **Figure 56**).

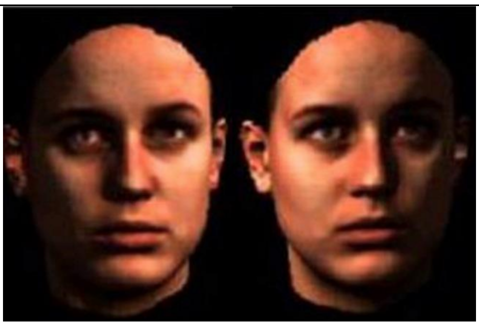


Figure 56 presents illuminated faces with direct light applied only to one side of the face. After showing the image on the left, participants were very accurate in judging if the face images on the right and left define the same person (Image reproduced from Braje et al. 1998).

The participants were very good at judging if two images representing the same face. Research results showed that direct light applied to face representations helped participants effectively matching faces even in the new unseen lighting conditions. However, when two faces were presented in diverse lighting conditions, it was much harder for the participants to decide whether these faces belong to one person than if the two faces were presented in the same lighting conditions. It shows that various lighting conditions have a different impact on face generalisation processes and indicates that matching faces in the novel lighting conditions is not so easy even for the human vision. Also, a combination of changing pose, the direction of a light source together with a complex structure of a human face can create a different shading and shadows on the face that will make a face hard to recognise. **Figure 57** shows examples of faces that were captured under varying illumination conditions and are hard to identify as belonging to the same person. Adini, Moses and Ullman (1997) proved that such variations in facial appearance could be a lot bigger than variations caused by changes in personal identity.

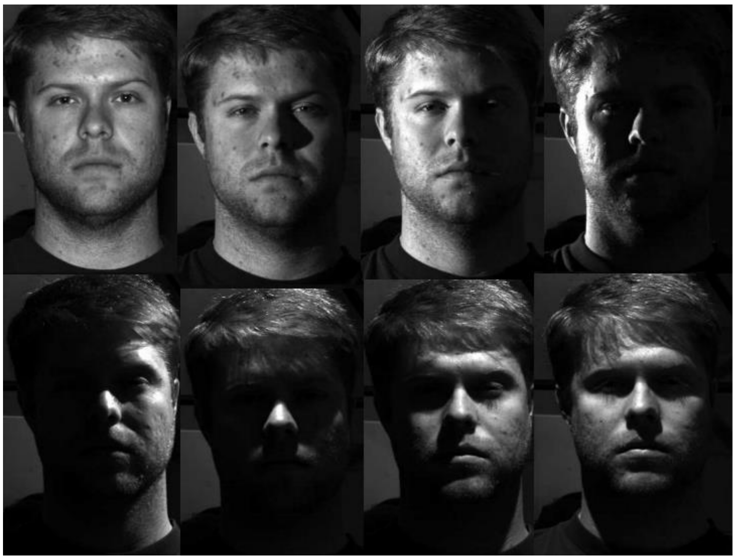


Figure 57 Yale Face Database B: examples of faces that are hard to recognise photographed under various lighting conditions. (taken from Yale database http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html)

From Chapter 1, we know that human vision is lighting dependent, on the other hand, humans evolved very effective and efficient mechanisms for the perception of faces in changing lighting conditions. One of the most important mechanisms that the human brain uses to recover the shape of the face is shape-from-shading. Shading is an essential component for humans to perceive the shape of the face. Images often illustrate subtle luminance variations (shading) as a function of spatial position. Hill and Bruce (1996) showed that the direction of light disrupts face matching performance. They showed that participants could not decide whether two face images represent the same person when a face in one frame was illuminated from above and another from below. Johnston et al. (1992) also found that even familiar faces were hard to recognise when they were illuminated from below rather than from above (see **Figure 58**).

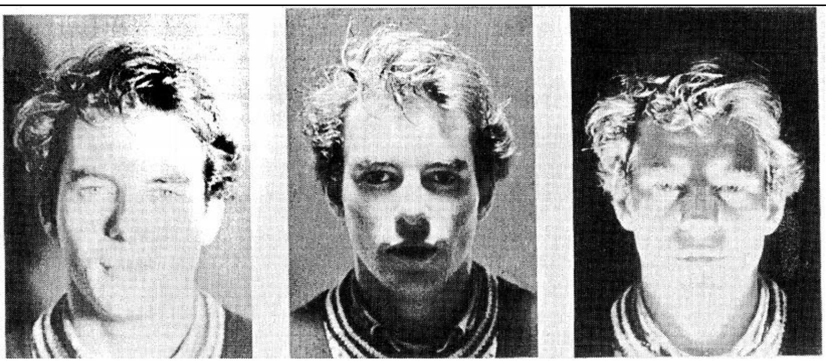


Figure 58 The images of the face photographed under different lighting conditions were used in Johnston et al.'s (1992) experiment to investigate if the images illustrate the same face. The image on the right represents a "ghost-like" face appearance which is harder to recognise compared to the rest of the image in the same row but results in a reduced inversion effect (image reproduced from Johnston et al., 1992).

In computer vision, shape-from-shading techniques are also candidate methods used for reconstructing a shape from shading in facial images, but shape-from-shading is a difficult challenge.

To construct an illuminant invariant descriptor, several steps have to be accomplished. First, in computer vision, the extraction of scale-invariant interest regions from colour images frequently begins with the conversion of the image to a grayscale image. Identification of interest points is then completely determined by luminance, and the use of colour is deferred to the stage of descriptor formation. Next, after obtaining an invariant image, shadows need to be removed. Removing shadows is usually done by finding edges of the objects in the images (Canny 1986; Lowe 2004). The shadowless image would be an image where large patches of shadow are removed, effectively taking out the direction of the illumination and leaving only the reflectance and the locally shaded texture.

Would this computer vision approach work with an extreme example, for instance, with Mooney face images? Mooney face images look like if they were illuminated with direct light from one side, giving a strong lightening effect on one side of the face and shadow on the other. The edges of the contours of these binary images are not directly related to the actual face contours. (**Figure 59**).

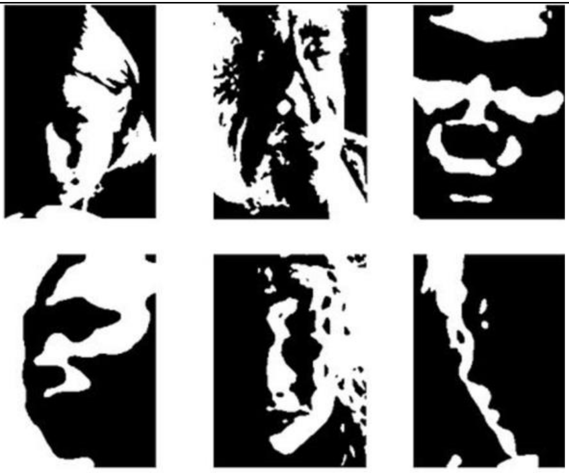


Figure 59 This figure illustrates Mooney faces used by Craig Mooney (Mooney, 1957) to test the ability of children to perform perceptual closure. Notice the variety of shapes and contours that emerge (image reproduced diplopi.wordpress.com. Reprinted 17 February 2020, from **https://diplopi.files.wordpress.com/2013/09/mooney_01.jpg**).

The standard computer vision strategy to overcome the lighting problem is to try to eliminate it by working with the shape of edges recovered by standard edge detection techniques, and thereby to remove shading from the face will not work on all cases, and Mooney's faces are a prime example of such a case. Therefore, a computer vision strategy would not work.

Thus this chapter proposes an alternative approach that will include all the shading information of the face to achieve face constancy under various lighting conditions, the same as the human vision system. We will explore whether the methods we used to generate invariance over the pose can also work for lighting.

5.1. Test stimuli: the database of multiple face sequences in different lighting.

To test this approach, four sets of facial motion sequences were created. The sequences were captured using a purpose-built illumination rig as described in 3.1 and 3.1.5. Each sequence is representing a non-rigid capture of facial behaviour in different illuminant conditions for one head pose. Sequences were captured using six Grasshopper GRAS-03K2C (FireWire) digital cameras, at a rate of 60 frames per second, at 640x480 image resolution with an RGB 24-bit pixel format. The sequences were of an expressive talking face, and they lasted around 30 seconds. However, we have chosen to analyse captured video sequence from only one Grasshopper camera, focusing on the illumination problem. We wanted to record dynamic facial sequences that were identical apart from differences in the direction of the illumination. To achieve this, the faces were illuminated simultaneously using red, green and blue lights projected from three projectors located in different positions. We then extracted the individual red, green and blue channels from the RGB image sequences and converted them into greyscale image sequences. This process gave us a synchronised data for multiple lighting directions. Also, there were four conditions of diffuse white lighting, which filled in areas without direct lighting. These lights were:

1. A diffuse light from the left, covering the left side of the participant's face.
2. A diffuse light from the right, covering the right side of the participant's face.
3. A diffuse light from the bottom, covering the bottom of the participant's face.
4. A diffuse light from the front, distributed symmetrically upon the participant's face.

Four different experiments were created, one for each of the diffuse lighting conditions. The motion sequences were slightly different in each experiment.

5.2. Experiment 1

The sequences were vectorised as described in 3.3.1. The sequences for each face were processed using PCA, and the basis vectors were extracted as described to form a PCA-based model of the target face.

5.2.1. First illuminance condition

In the first experiment, we used faces that were diffusely illuminated from the left and directly illuminated by three coloured lights from different directions to test whether we could recover an estimate of how the face would look if illuminated from one of the other directions. The results of the first experiment show that the missing illuminated faces can be correctly reconstructed using partial PCA in which the probe data is drawn from just one illuminated face, as can be seen in **Figure 60**. First, we constructed a multi-vector PCA space where each input vector to the PCA contained data from all three illuminants (plus the ambient lighting) captured at each time point. This PCA face-space was created to hold the images generated by the three main illuminants within a content addressable memory (Kohonen, 1986). To reconstruct the missing illuminated faces from just one illuminated face, we mapped an image vector derived from that illuminated face into our multi-illuminant content addressable memory.

Ambient light source covering the left side of the participant’s face

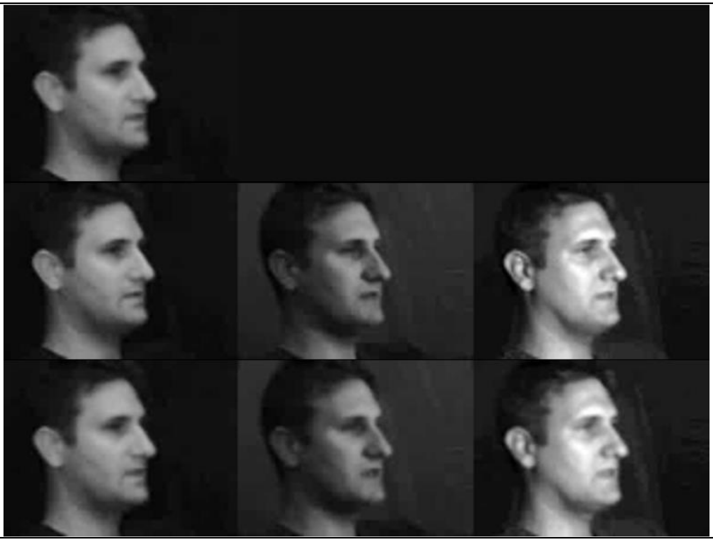
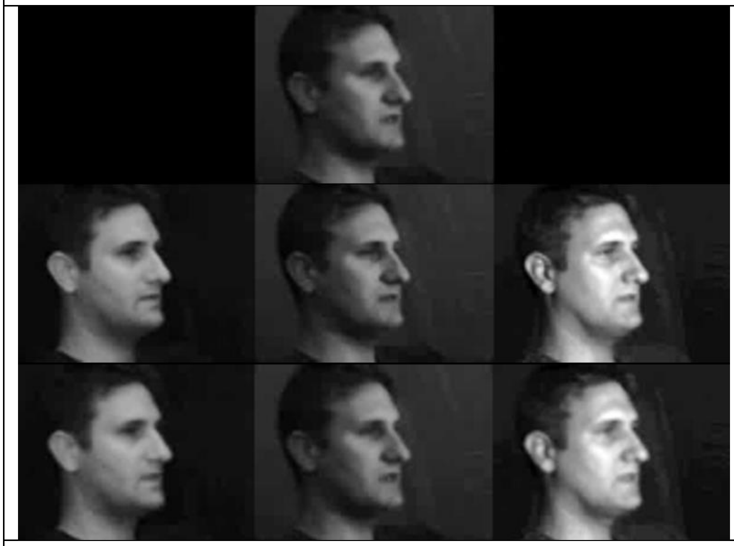
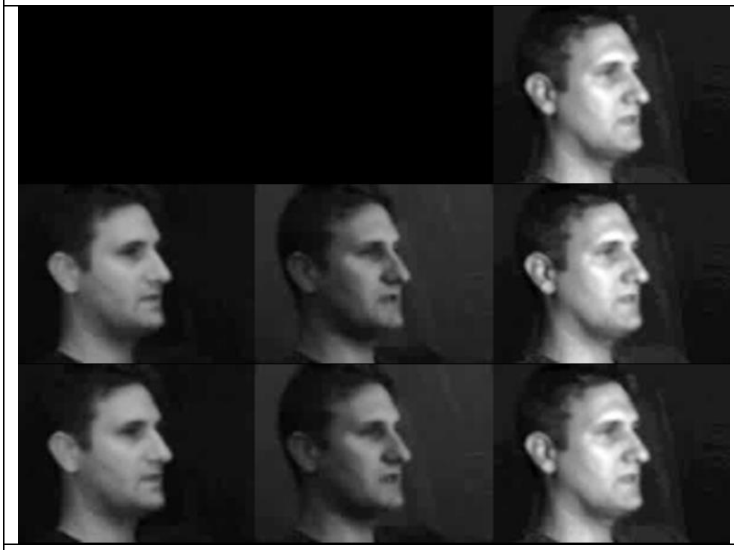
1st Illuminant	
	The 1st illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 1st illuminant image information.
	The ground-truth multi-illuminant image representation.
2nd Illuminant	
	The 2nd directional illuminant data in the multi-illuminant image is present. The rest of the information is taken away (blacked out).
	The multi-illuminant image reconstructed using only the 2nd directional illuminant data image information.
	The ground-truth multi-illuminant image representation.
3rd Illuminant	
	The 3rd directional illuminant data in the multi-illuminant image is present. The rest of the information is taken away (blacked out).
	The multi-illuminant image reconstructed using only the 3rd directional illuminant data image information.
	The ground-truth multi-illuminant image representation.

Figure 60 The *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* images show the reconstruction of different facial expressions from the information of the face derived solely from the *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* (see **Movie 4**.avi).

The multi-illuminant images represented in **Figure 60** are results of the facial reconstruction from single illuminant images. From visual inspection reconstructions shown in **Figure 60,** they appear identical to the ground-truth. However, if we take a closer look at the reconstructed results of all illuminants, we may notice that the mouth of the reconstructed image is slightly closed compared to the same ground-truth expression. However, the reconstructed mouth in the 3rd illuminant condition provided a better match to the ground-truth than the rest. Our observations are confirmed in **Figure 61**. There is a better match between the ground-truth data and the reconstructed data from 3rd illuminant; attempts to recover the other two illuminant images were less effective. However, we are looking to find the best match with ground-truth data representation, hence we will find a scaling factor.

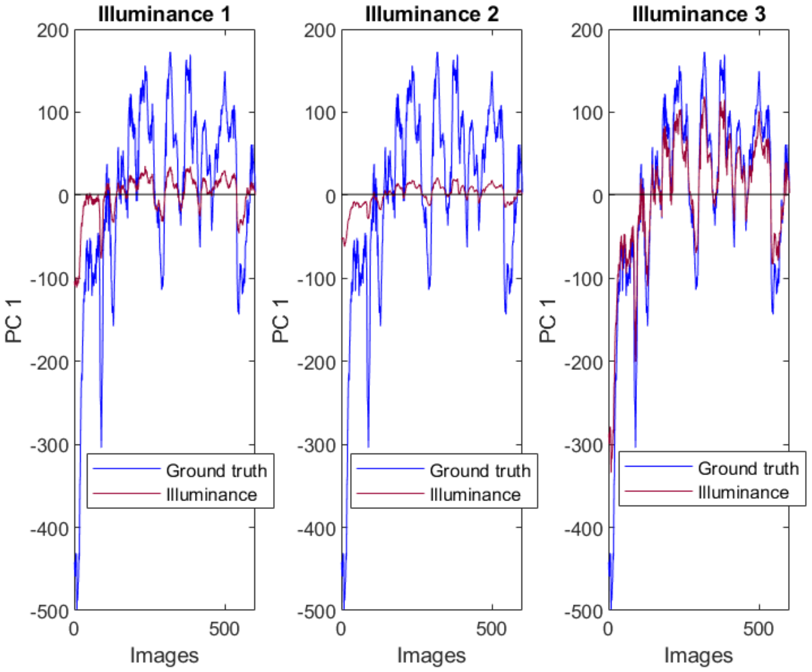


Figure 61 All three graphs represent loadings comparisons for reconstructed multi-illuminant images compared to the ground-truth loadings. The graph on the left shows loadings from the 1st illuminant (red colour) are close to the ground-truth (blue colour) multi-illuminant images, but the values of the loadings are smaller overall. The middle graph shows a loadings time series for the reconstructed multi-illuminant data seeded using the 2nd illuminant (red colour) compared to the ground-truth (blue colour). The right graph presents the loadings time series for the reconstructed multi-illuminant images derived from the 3rd illuminant (red colour) compared to ground-truth multi-illuminant data (blue colour). We can see from the graphs that the loadings of the reconstructed multi-illuminant images from the 3rd illuminant are very close to the ground-truth multi-illuminant representation. That means that the images should be very similar to each other. The loadings of the other two graphs are less close to the ground-truth; hence the reconstructed images should contain less similarities with the ground-truth images.

We used a linear least-squares method to find a scale factor that generated the best correspondences between a particular illuminant and the ground-truth loadings time series. A scaling factor will improve a data representation which will give more coherent facial movements with presented delicate non-rigid changes of a face that were barely noticeable in reconstructed sequences before applying the scaling coefficient. The scaling factor for every principal component (ten principal components) over three illuminants are shown in **Figure 62.**

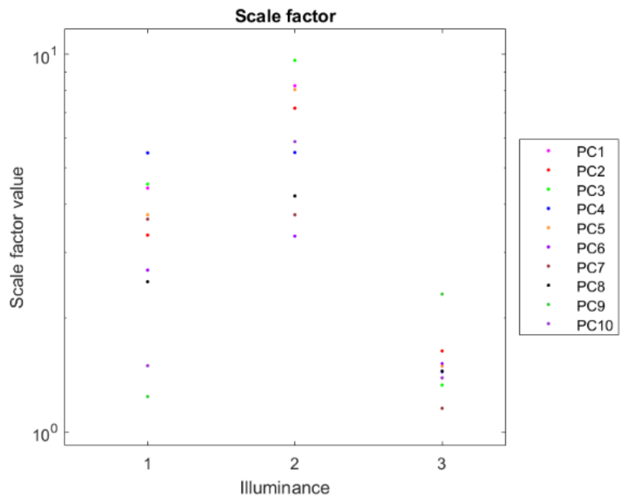


Figure 62 The figure presents a scaling factor data distribution for the 1st, 2nd and 3rd illuminant. As we may see, the largest data range is for 1st illuminant. The 2nd illuminant scaling factor data range is smaller than the 1st illuminant's. The 3rd illuminant has the smallest scaling factor.

From **Figure 62,** we may see that for the 3rd illuminant reconstruction, the scaling factor is much smaller than the scaling factor for the 1st and 2nd illuminant. That is because the 1st and 2nd illuminants are darker (contain less information) than the brighter 3rd illuminant. Hence their scaling factor will require a larger value to be able to match ground-truth.

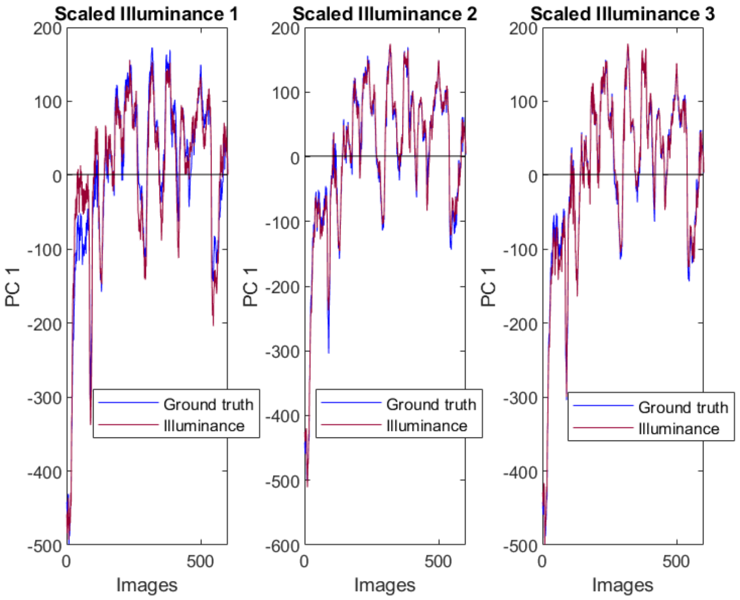
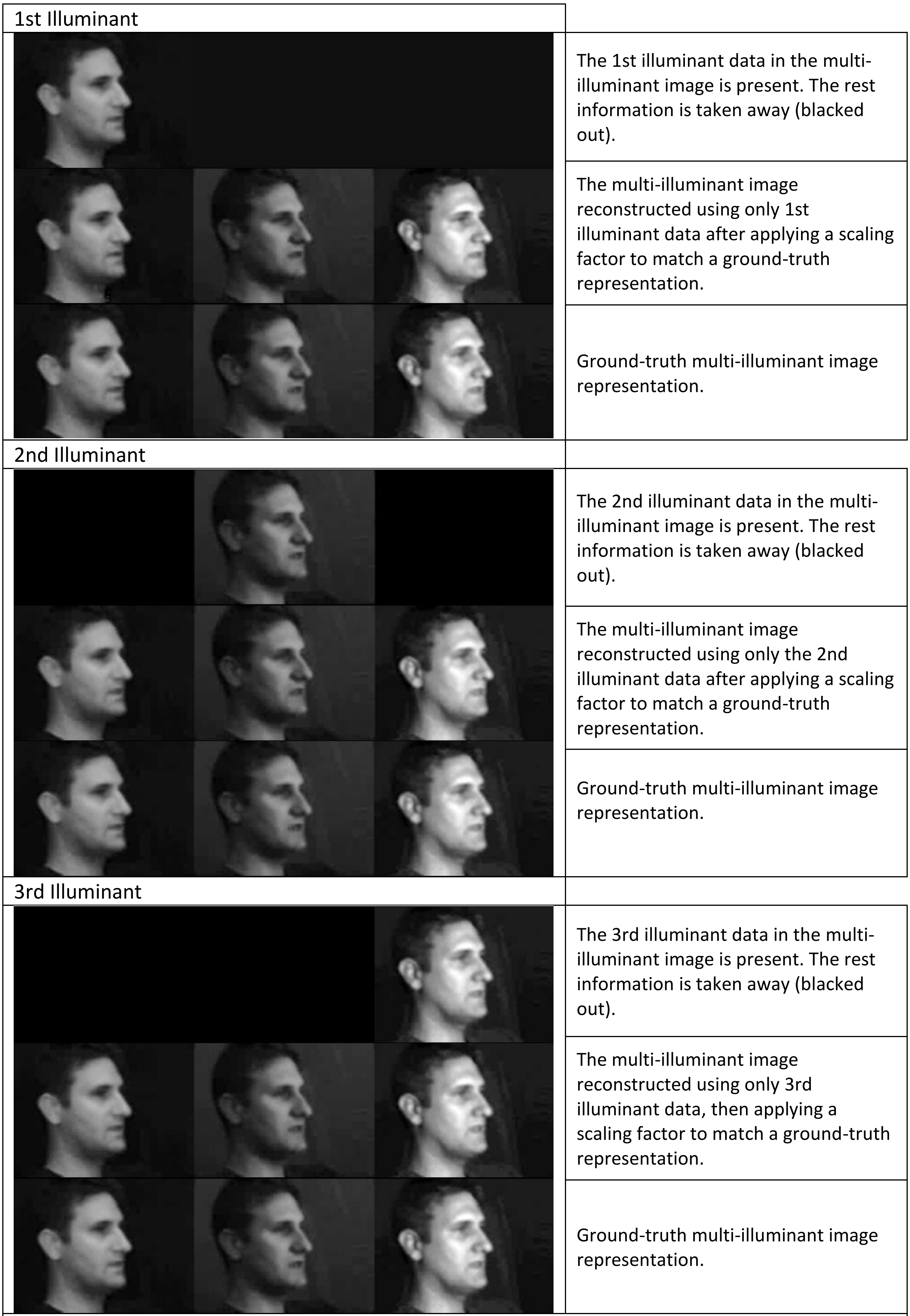


Figure 63 Loadings after scaling for the 1st, 2nd and 3rd illuminant (red colour) compared to ground-truth loadings (blue colour). The scaled loadings provide a better match to the ground-truth.

In **Figure 63**, we can see a better match between the scaled reconstructed results and the ground-truth image representation. The results of reconstructing images after applying the scaling factor are presented in **Figure 64**.



1st Illuminant

The 1st illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).

The multi-illuminant image reconstructed using only 1st illuminant data after applying a scaling factor to match a ground-truth representation.

Ground-truth multi-illuminant image representation.

2nd Illuminant

The 2nd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).

The multi-illuminant image reconstructed using only the 2nd illuminant data after applying a scaling factor to match a ground-truth representation.

Ground-truth multi-illuminant image representation.

3rd Illuminant

The 3rd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).

The multi-illuminant image reconstructed using only 3rd illuminant data, then applying a scaling factor to match a ground-truth representation.

Ground-truth multi-illuminant image representation.

Figure 64 *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* show reconstructions of facial expressions having only information of the face from the *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant,* respectively (see Movie 5.avi).

From a visual inspection of **Figure 64,** it can be seen that the reconstructions for all illuminants after applying a scaling factor to our content addressable memory technique look very accurate and close to the ground-truth representation. This improvement in accuracy is to be expected because we obtained the scaling factor explicitly for each principal component and illuminant.

To measure how close our scaled reconstructed multi-illuminant images are to the ground-truth images, we employed a standard Pearson product-moment correlation coefficient approach (PPMCC) (Stigler 1989). The PPMCC finds a linear dependence between two variables by giving a value between +1 and -1 inclusively, where 1 indicates a perfect positive linear correlation, 0 is no linear correlation, -1 indicates a perfect negative correlation. This approach was used to define the degree to which the scaled reconstructed illuminated image is close to the original illuminated image representation. **Figure 65** shows the image correlation between principal component weightings extracted from the 600 black out at a specific part of multi-view driver sequence vectors and of those from the ground-truth. We may observe in the three histograms of the right column how correlation values are distributed for every illuminant. We may notice that a range of correlation values differs for every illuminant. Values for the 1st illuminant are ranging from 0.4159 to 0.9408, for the 2nd illuminant from 0.498 to 0.9956, and for the 3rd illuminant from 0.7541 to 0.9984. Also, we may observe that the number of facial expressions with the correlation coefficient for all three illuminants in the histogram graphs mainly exceeds 0.8. For this reason, we have found the percentage value that shows how much data possesses a correlation coefficient above 0.8. 1st illuminant had 91,5 per cent, 2nd illuminant had 93,83 per cent, and 3rd illuminant had 99.66 per cent from of all data. In the graphs of the left columns, we have investigated correlation coefficients between the reconstructed multi-illuminant image from 1st, 2nd, and 3rd illuminants and ground-truth accordingly of a randomly chosen image from the sequence. The correlation values were ranging from 0.9908 and 0.9888. The highest correlation was obtained from 3rd illuminant with a value of 0.9888. Then 2nd illuminant with value 0.9833, and 3rd illuminant with value 0.9408. It means that reconstructed weights of images are highly correlated, i.e. the reconstructed multi-illuminant images are very close to the ground-truth. The high correlations are not surprising given these are image correlations between very similar images, so in order to provide a relative test of the quality of reconstruction, we used the same analysis method described in Chapter 4 for testing the quality of pose reconstruction.

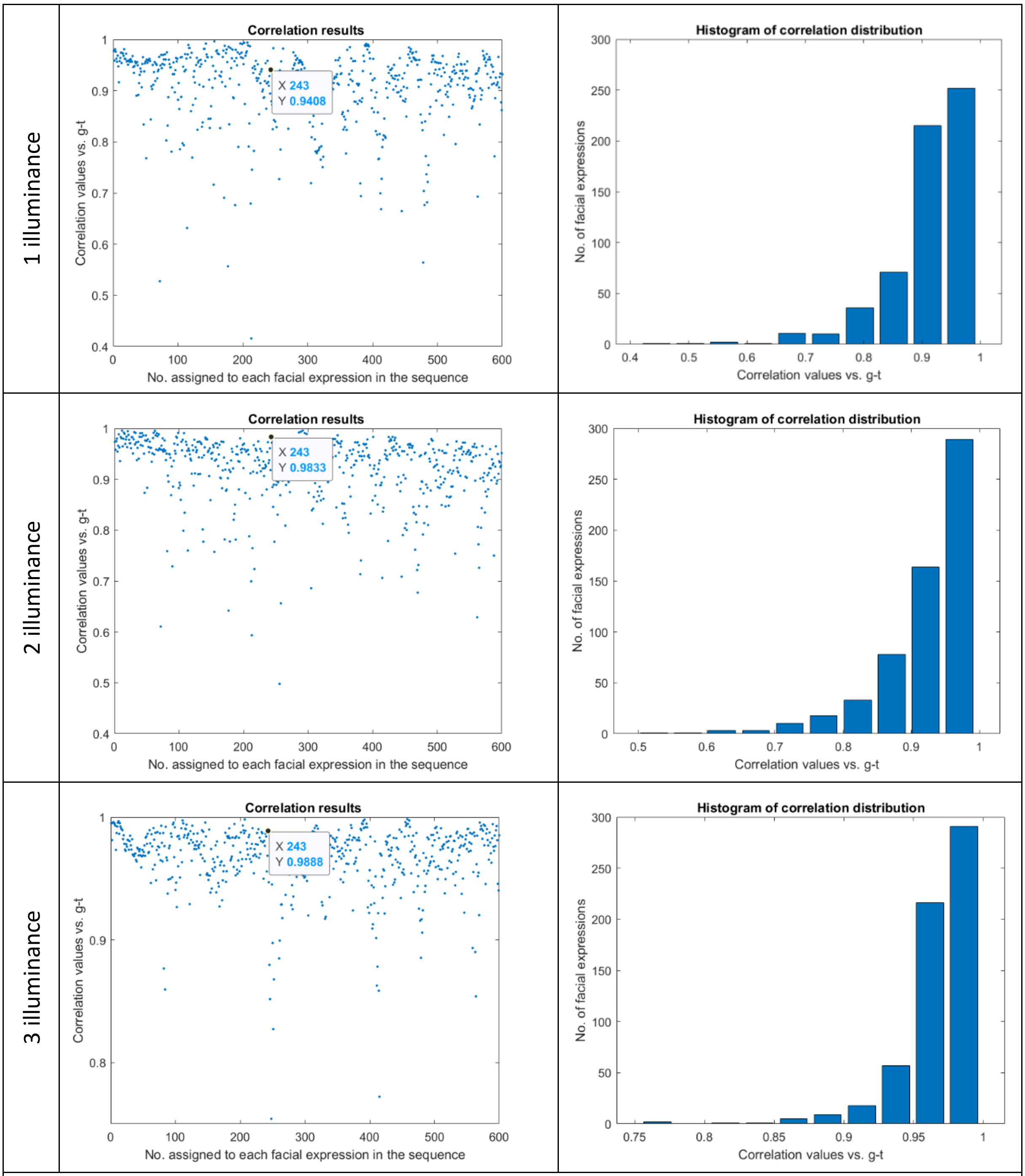


Figure 65 shows the correlation results. The points in three graphs of the left column show PC scores relation that defines how close the scaled/reconstructed 1st illuminant, 2nd illuminant, and 3rd illuminant are to the ground-truth. Also, a point in these graphs, at position 243 shows the correlation between the scaled/reconstructed 1st, 2nd and 3rd illuminants and the ground-truth accordingly. The blue bars in three graphs of the right column show the distribution of correlation between pairs of image weights of reconstructed illuminants and the ground-truth data.

5.2.2. Second illuminance condition

For the second experiment, we used the data from the faces illuminated from the right as input into the reconstruction process and repeated the steps used in the first experiment. The results of the second experiment showed that the missing illuminated faces could be correctly reconstructed by the system from only one illuminated face, as it can be observed in **Figure 66.**

Light source covering the right side of participant’s face

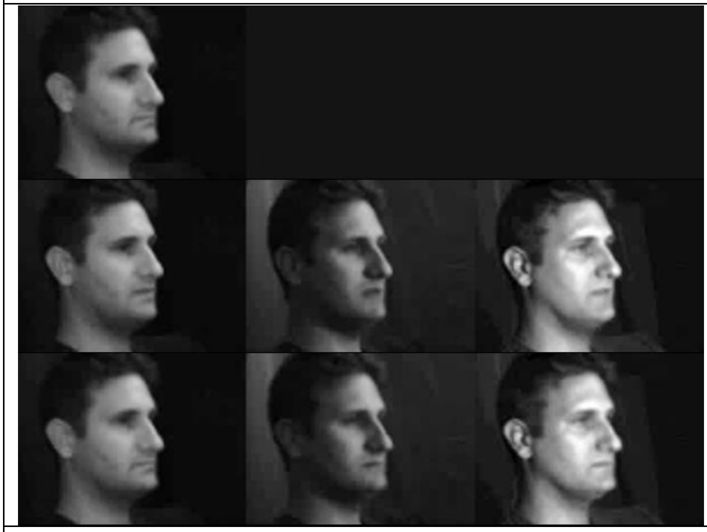
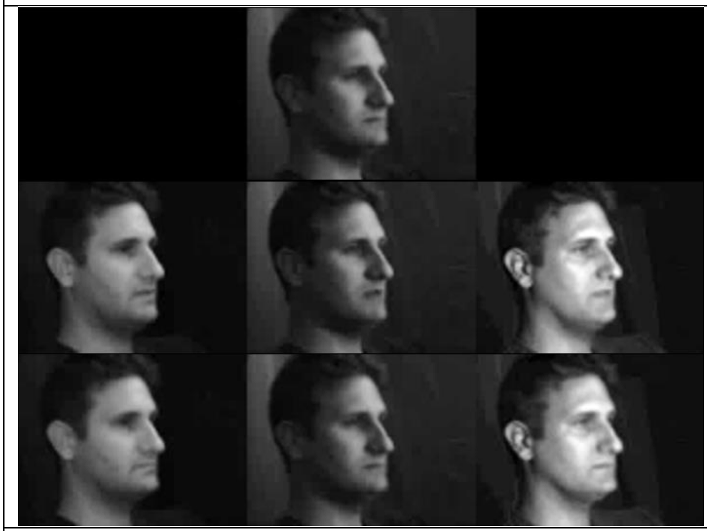
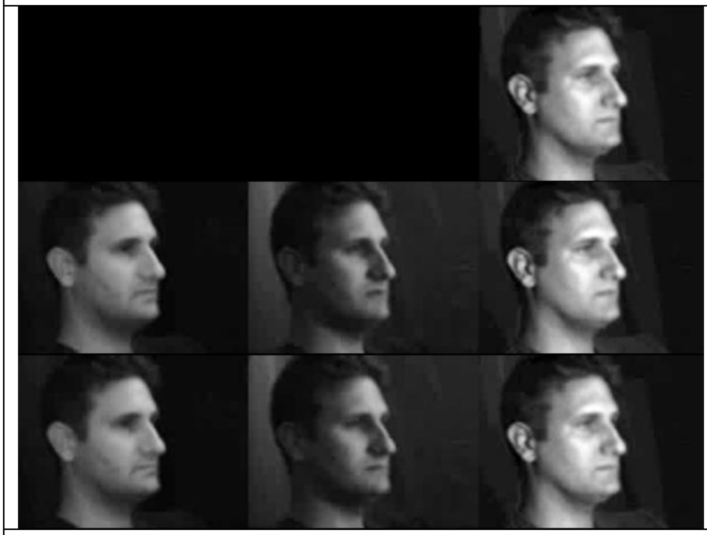
1st Illuminant	
	The 1st illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 1st illuminant data image information.
	Ground-truth multi-illuminant image representation.
2nd Illuminant	
	The 2nd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only the 2nd illuminant data image information.
	Ground-truth multi-illuminant image representation.
3rd Illuminant	
	The 3rd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 3rd illuminant data image information.
	Ground-truth multi-illuminant image representation.

Figure 66 ***1st Illuminant***, ***2nd Illuminant*** and ***3rd Illuminant*** show reconstruction of different facial expressions having only information of the face from ***1st Illuminant***, ***2nd Illuminant*** and ***3rd Illuminant*** (see **Movie 6**.avi).

The multi-illuminant images shown in **Figure 66** are the results of the facial reconstruction using different illuminants while the diffuse light was covering the left side of the face. From visual inspection of the reconstructions shown in **Figure 66,** they appear very similar to the ground-truth. Let us take a closer look at the reconstructed multi-illuminant frame of 1st and 2nd illuminant; we can see that the mouth of the participant is slightly open compared to the ground-truth participant with the same expression. However, the reconstructed mouth of a participant in the 3rd illuminant gave a better match with ground-truth than the rest. These results are very similar to the results from the first illuminance condition.

From **Figure 67,** we may see our observations clearly. The ground-truth data have a better match with reconstructed data from the 3rd illuminant; the other two remain similar with a small match with ground-truth. However, we are looking to find the best match with ground-truth data representation; hence again, we will find a scaling factor.

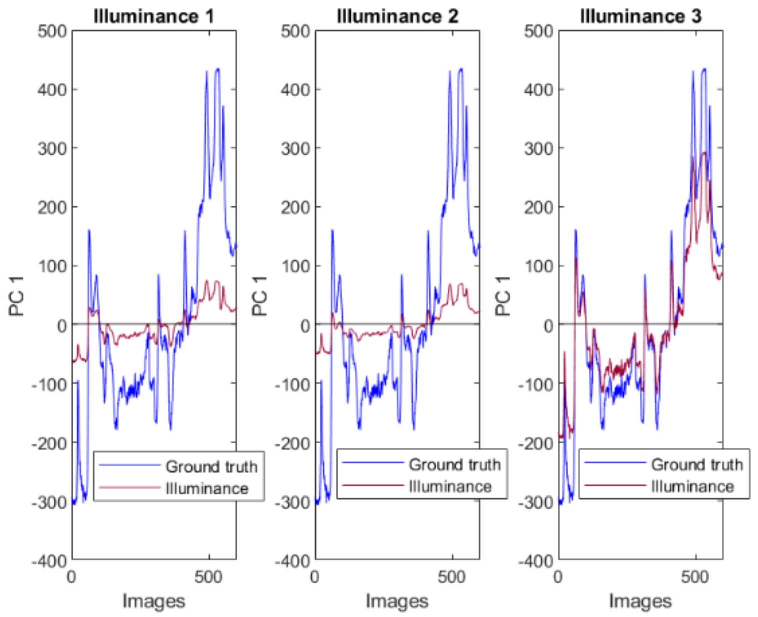


Figure 67 all three graphs present loadings comparison of reconstructed multi-illuminant images with ground-truth representation. The graph on the left shows how loadings of reconstructed multi-illuminant images from 1st illuminant (red colour) are close to the ground-truth (blue colour) multi-illuminant images. The middle graph shows a loadings distribution for reconstructed multi-illuminant from 2nd illuminant (red colour) and ground-truth (blue colour). The right graph presents loadings distribution of reconstructed multi-illuminant images from the 3rd illuminant (red colour) with ground-truth multi-illuminant images (blue colour). As we may notice from the graphs, the loadings of reconstructed multi-illuminant images from the 3rd illuminant are very close to the ground-truth multi-illuminant representation; it means that images should be very similar to each other. The loadings of the other two graphs are less close to the ground-truth, hence reconstructed images should contain fewer similarities with ground-truth images.

Next, we found a scaling factor as before. From **Figure 68**, we can see that, as already observed, the scale factor range is much smaller for the 1st and 2nd illuminant for the same reason that the 3rd illuminant is brighter.

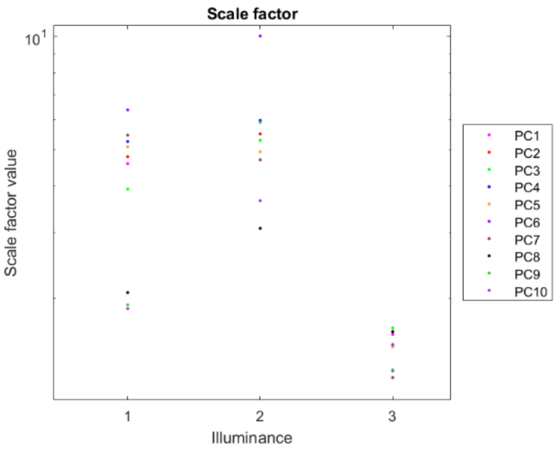


Figure 68 The figure presents a scaling factor data distribution for the 1st, 2nd and 3rd illuminant. As we may see, the largest data range is for 1st illuminant. The 2nd illuminant scaling factor data range is smaller than the 1st illuminant. The 3rd illuminant has the smallest scaling factor.

Figure 68 shows that the scaling factor for the 3rd illuminant reconstructed view is much smaller than the scaling factor for the 1st and 2nd illuminant. It is due to 1st, and 2nd illuminant being darker (contain less information) than (bright) the 3rd illuminant. Hence scaling factor of 1st and 2nd illuminants will have to be bigger to be able to match ground-truth.

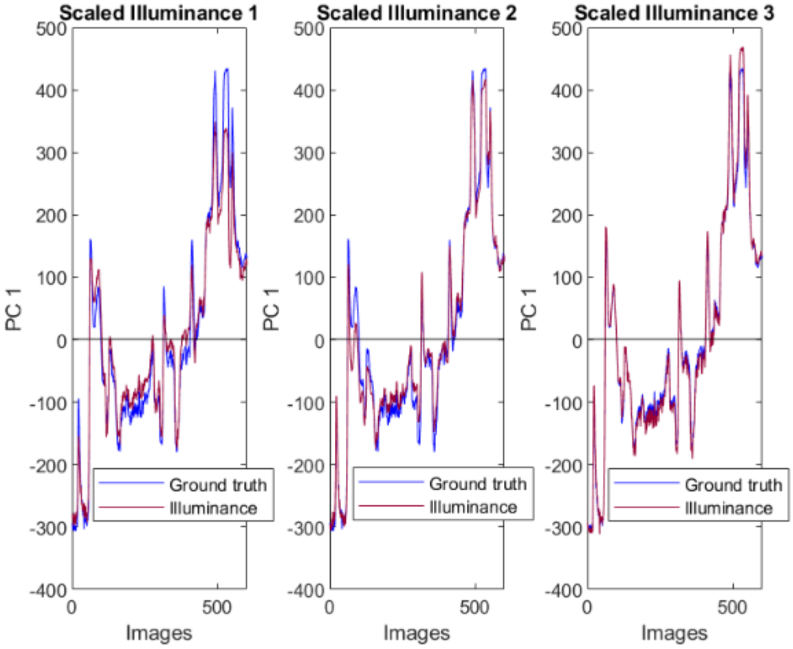
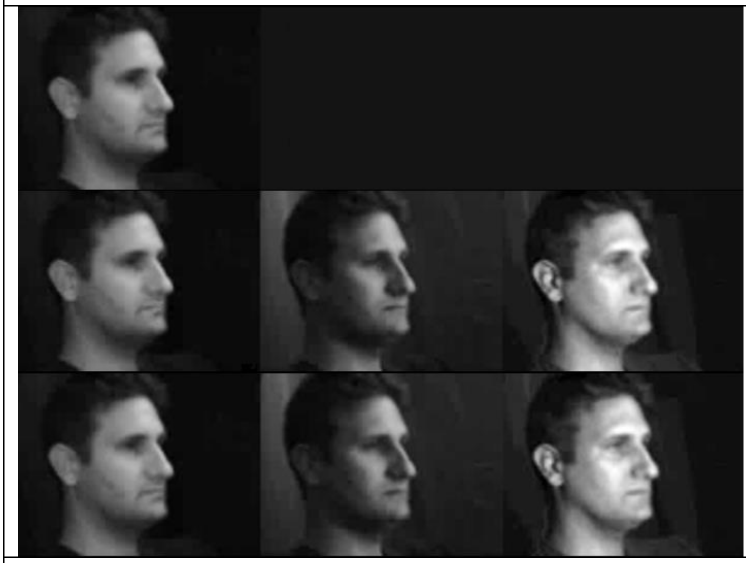


Figure 69 The graph presents loadings matching of scaled loadings from 1st, 2nd and 3rd illuminant (red colour) with ground-truth loadings (blue colour). As we may notice from the graphs, scaled loadings gave a better match to ground-truth.

In **Figure 69,** we can observe a better match of scaled reconstructed results with ground-truth image representation. The results of reconstructed images after applying a scaling factor are presented in **Figure 70**.

1st Illuminant

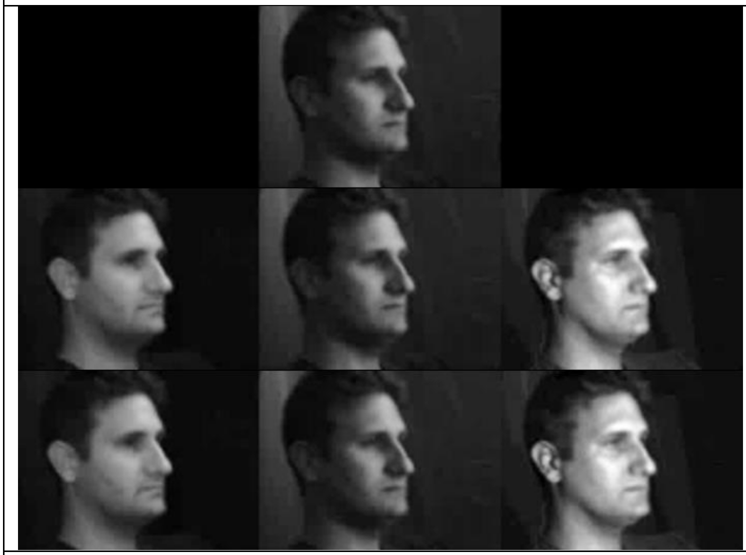


The 1st illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).

The multi-illuminant image reconstructed using only 1st illuminant data then applying a scaling factor to match a ground-truth representation.

Ground-truth multi-illuminant image representation.

2nd Illuminant

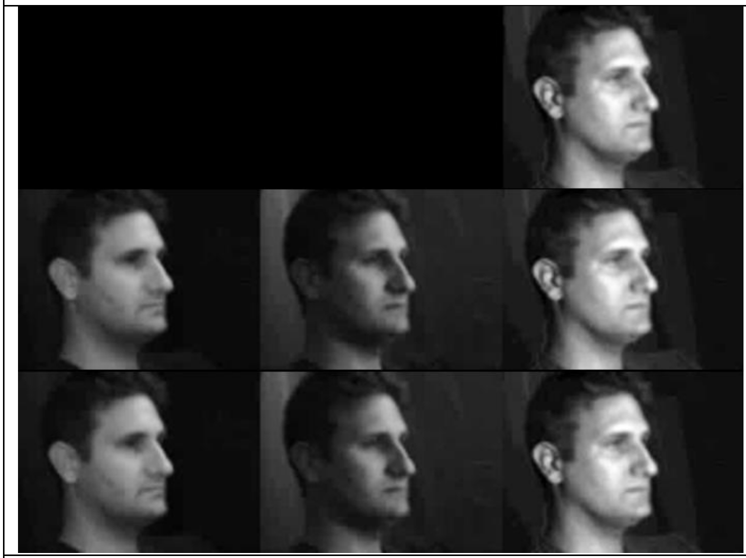


The 2nd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).

The multi-illuminant image reconstructed using only the 2nd illuminant data after applying a scaling factor to match a ground-truth representation.

Ground-truth multi-illuminant image representation.

3rd Illuminant



The 3rd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).

The multi-illuminant image reconstructed using only 3rd illuminant data after applying a scaling factor to match a ground-truth representation.

Ground-truth multi-illuminant image representation.

Figure 70 *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* show reconstruction of different facial expressions having only information of the face from *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* (see **Movie 7**.avi).

From visual inspection in **Figure 70,** it can be seen that the scaled reconstructed image for all illuminants looks very accurate and close to the ground-truth representation in our content addressable memory. The accuracy was to be expected because the scaling factor has been obtained explicitly for each principal component and illuminant.

Comparing the correlation coefficients of facial expression at position 281 in graphs of the left column in **Figure 71,** we see a very high correlation value between the weight of scaled/reconstructed data and ground-truth. As in the first experiment, the correlations are close to 1. That means that the scaled reconstructed images from multiple illuminants are a close match to original data representation. In addition, we may observe in the histogram graphs of the right column how correlation data is distributed for all illuminants.

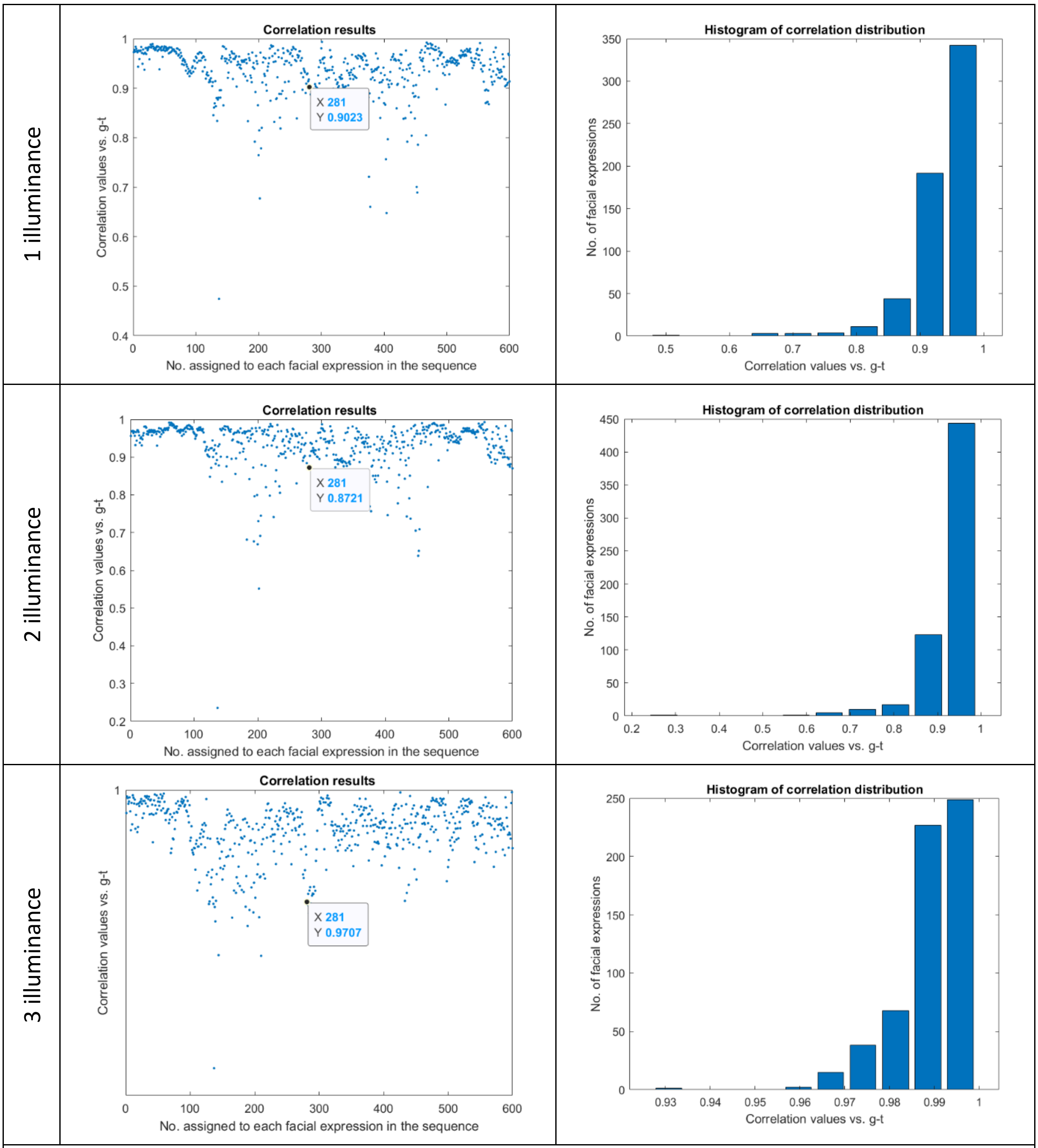


Figure 71 shows the correlation results. The points in three graphs of the left column show PC scores relation that defines how close the scaled/reconstructed 1st illuminant, 2nd illuminant, and 3rd illuminant are to the ground-truth. Also, a point in these graphs, at position 281 shows the correlation between the scaled/reconstructed 1st, 2nd and 3rd illuminants and the ground-truth accordingly. The blue bars in three graphs of the right column show the distribution of correlation between pairs of image weights of reconstructed illuminants and the ground-truth data.

5.2.3. Third illuminance condition

For third illuminance condition, controlled lights have been used with diffuse white light highlighting the bottom part of the face and applied the same analysis introduced in first illuminance condition to a third data set. The results of the third illuminance condition showed that the missing illuminated faces could be correctly reconstructed from only one illuminated face, see in **Figure 72**.

Light source covering the bottom side of participant’s face

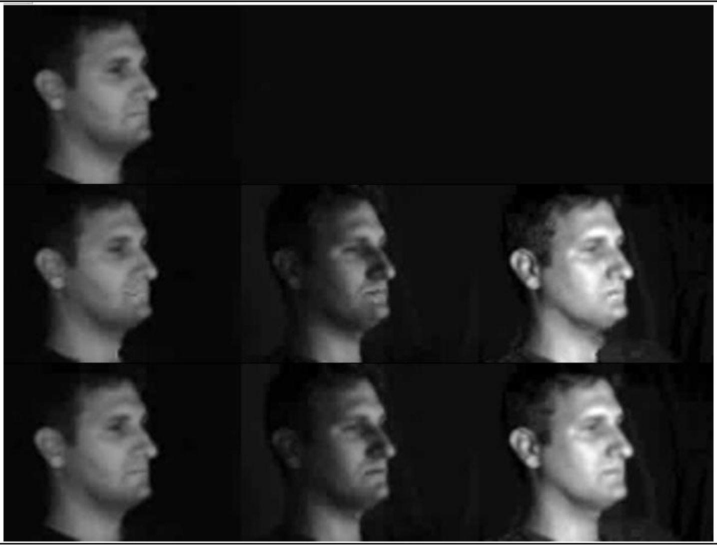
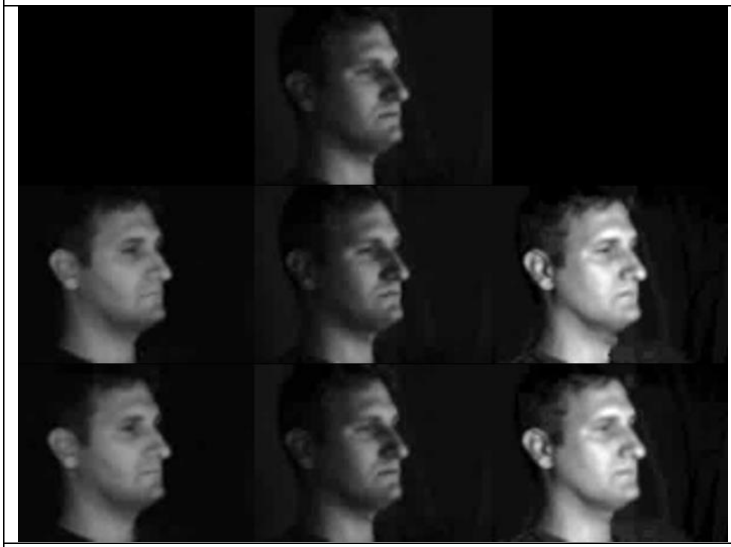
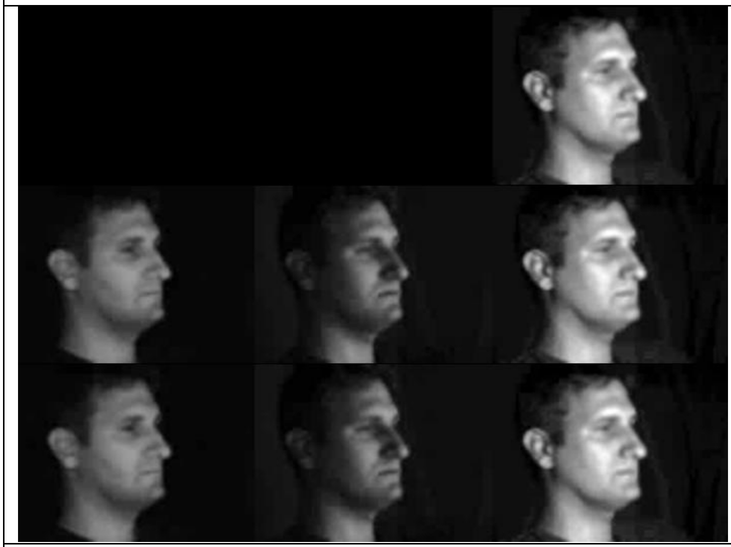
1st Illuminant	
	The 1st illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 1st illuminant data image information.
	Ground-truth multi-illuminant image representation.
2nd Illuminant	
	The 2nd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only the 2nd illuminant data image information.
	Ground-truth multi-illuminant image representation.
3rd Illuminant	
	The 3rd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 3rd illuminant data image information.
	Ground-truth multi-illuminant image representation.

Figure 72 *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* show reconstruction of different facial expressions having only information of the face from *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* (see Movie 8.avi).

The multi-illuminant images represented in **Figure 72** are the results of the facial reconstruction from a different illuminant. From visual inspection in **Figure 72,** reconstructions from the 3rd illuminant appear almost identical to ground-truth. However, reconstructions from the 1st and 2nd illuminant have very noticeable face defects around the mouth area. We did not see such obvious image differences in our previous experiments between the reconstructed and original data representation. **Figure 73** confirms the results we see with the naked eye. In **Figure 73,** the ground-truth data have a better match with the reconstructed data from the 3rd illuminant; the other two provide a very weak match with ground-truth. However, we are looking to find the best match with ground-truth data representation; hence again, we will find a scaling factor.

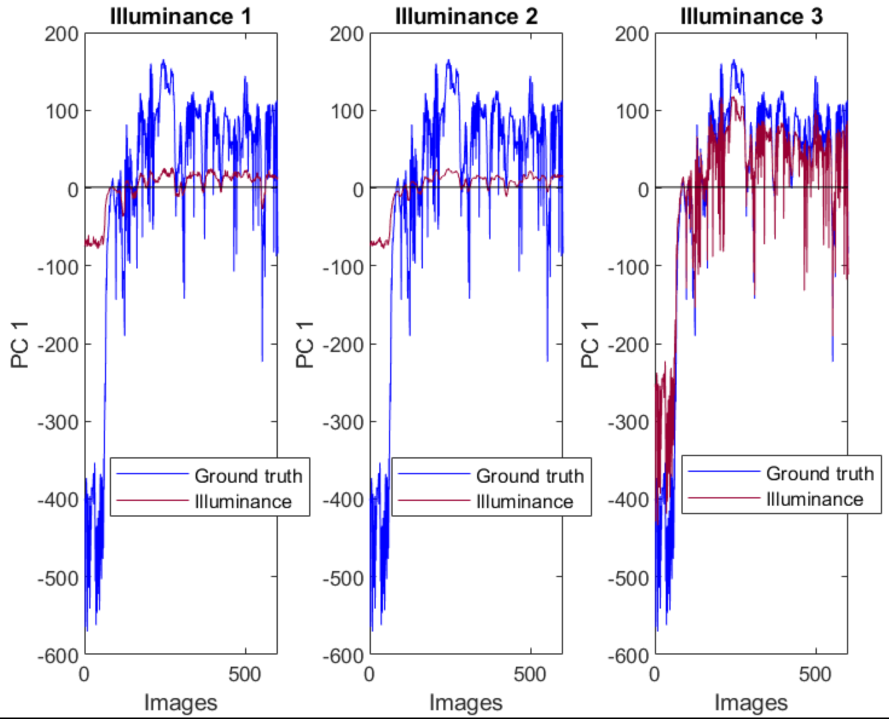


Figure 73 all three graphs represent loadings comparison of reconstructed multi-illuminant images with ground-truth representation. The graph on the left shows how loadings of reconstructed multi-illuminant images from 1st illuminant (red colour) are close to the ground-truth (blue colour) multi-illuminant images. The middle graph shows a loadings distribution for reconstructed multi-illuminant from 2nd illuminant (red colour) and ground-truth (blue colour). The right graph present loadings distribution of reconstructed multi-illuminant images from 3rd illuminant (red colour) with ground-truth multi-illuminant images (blue colour). As we may notice from the graphs, the loadings of reconstructed multi-illuminant images from the 3rd illuminant are very close to the ground-truth multi-illuminant representation; it means that images should be very similar to each other. The loadings of the other two graphs are less close to the ground-truth, hence reconstructed images should contain fewer similarities with ground-truth images.

From **Figure 74,** we may see that the scaling factor range for the reconstructed view of the 3rd illuminant is much smaller than the scaling factor range for the 1st and 2nd illuminant. It is due to the 1st and 2nd illuminant being darker than the 3rd illuminant image. Hence their scaling factor will need to have a bigger value to be able to match the ground-truth. We can see the same pattern in the scaling factor data as in the earlier experiments.

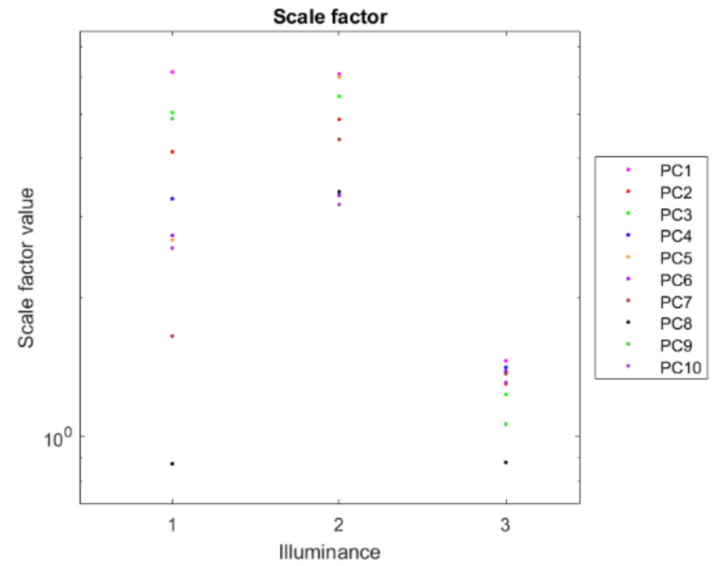


Figure 74 The figure presents a scaling factor data distribution for the 1st, 2nd and 3rd illuminant. As we may see, the largest data range is for 1st illuminant. The 2nd illuminant scaling factor data range is smaller than the 1st illuminant. The 3rd illuminant has the smallest scaling factor.

Figure 74 shows that the scaling factor of the 3rd illuminant reconstructed view is much smaller than the scaling factor for the 1st and 2nd illuminant. That is due to the 1st and 2nd illuminant being darker than the (bright) 3rd illuminant image. Hence the scaling factors will need to be bigger to be able to match a ground-truth.

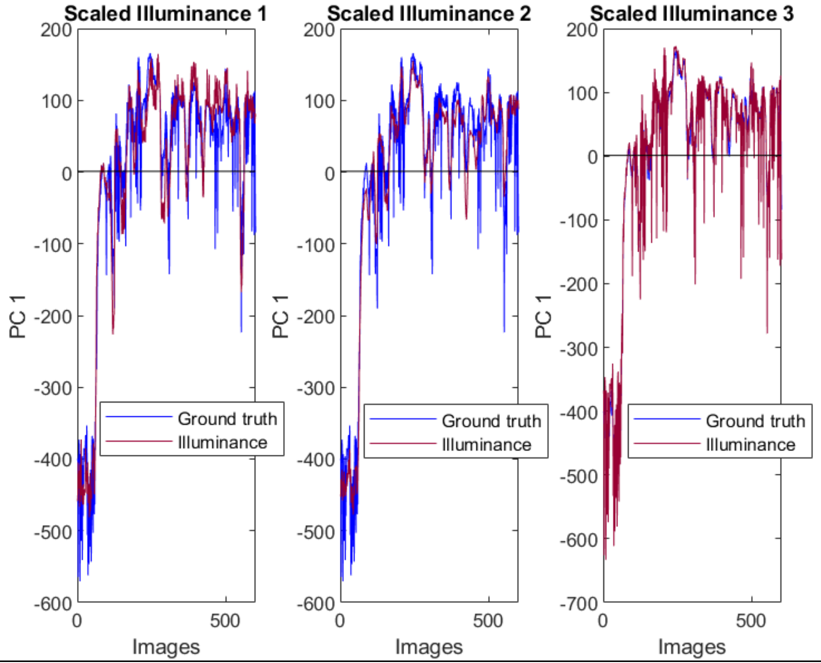
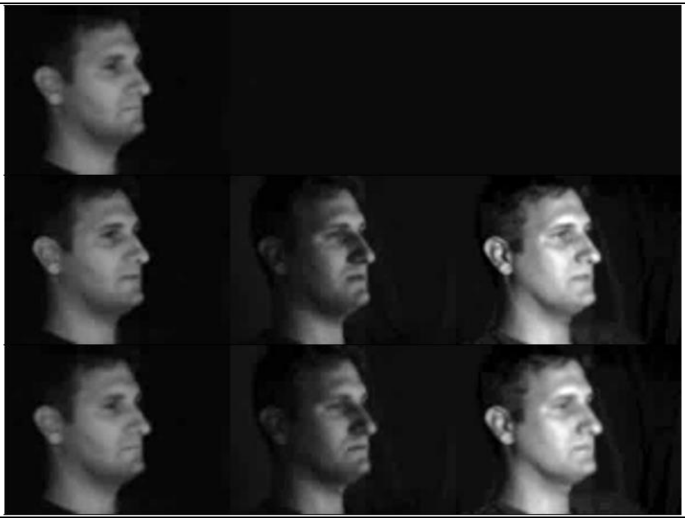
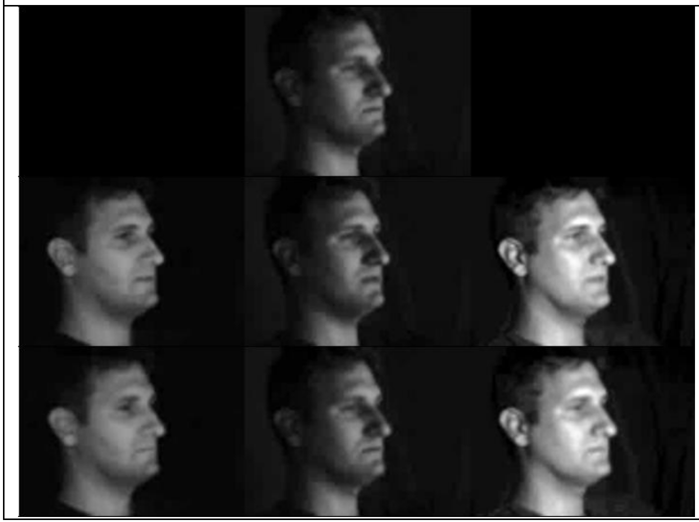


Figure 75 The graph presents loadings matching of scaled loadings from 1st 2nd and 3rd illuminant (red colour) with ground-truth loadings (blue colour). As we may notice from the graphs, scaled loadings gave a better matched to ground-truth.

In **Figure 75,** we may observe a better match of scaled reconstructed results with the ground-truth image representation. The results of reconstructed images after applying a scaling factor are presented in **Figure 76**.

Uniform lighting

1st Illuminant	
	The 1st illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out). The multi-illuminant image reconstructed using only 1st illuminant data then applying a scaling factor to match a ground-truth representation. Ground-truth multi-illuminant image representation.

2nd Illuminant	
	The 2nd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out). The multi-illuminant image reconstructed using only 2nd illuminant data after applying a scaling factor to match a ground-truth representation. Ground-truth multi-illuminant image representation.

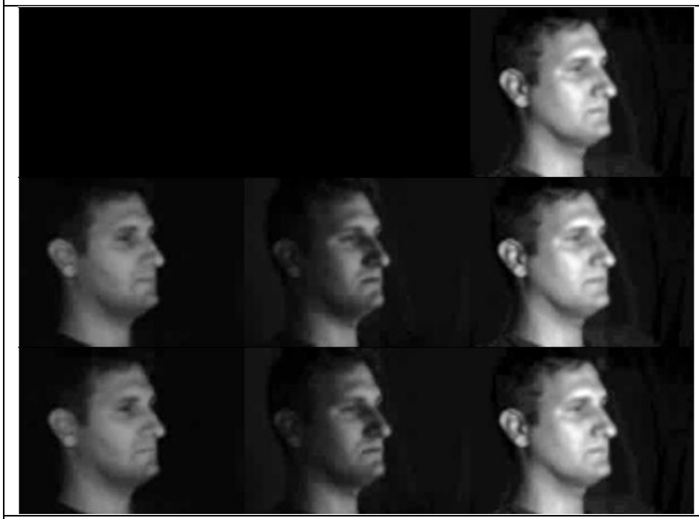
3rd Illuminant	
	The 3rd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out). The multi-illuminant image reconstructed using only 3rd illuminant data after applying a scaling factor to match a ground-truth representation. Ground-truth multi-illuminant image representation.

Figure 76 *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* show reconstruction of different facial expressions having only information of the face from *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* (see **Movie 9.**avi).

From visual inspection in **Figure 76,** we can see that the reconstructed image for all illuminants after applying a scaling factor looks very close to the ground-truth representation derived from our content addressable memory. The accuracy was to be expected because the scaling factors have been obtained individually for each principal component and illuminant. However, the system found it challenging to reconstruct the other illuminant conditions when the probe face was illuminated from below.

The correlation coefficients of randomly chosen facial expression at position 279 in graphs of the left side column in **Figure 77** shows a very high correlation value between the weights of scaled/reconstructed data and ground-truth. As in the first and second experiments, the correlation exceeds 0.8. That means that the reconstructed and the later scaled images derived from multiple illuminants are a close match to the original data representation. The histogram graphs of the right column show correlation distribution for every illuminant. As we may notice, a range of correlation values for 1st illuminant is between 0.2015 and 0.9791 when 2nd illuminant is in 0.1582 to 0.9885, and 3rd illuminant is in 0.6924 to 0.9982. Despite such a wide range of correlation values, we established the percentage of data that have a correlation coefficient above 0.8. 1st illuminant had 52.2 per cent of all data, 2nd illuminant had 67.2 per cent, and 3rd illuminant had 94.3 per cent.

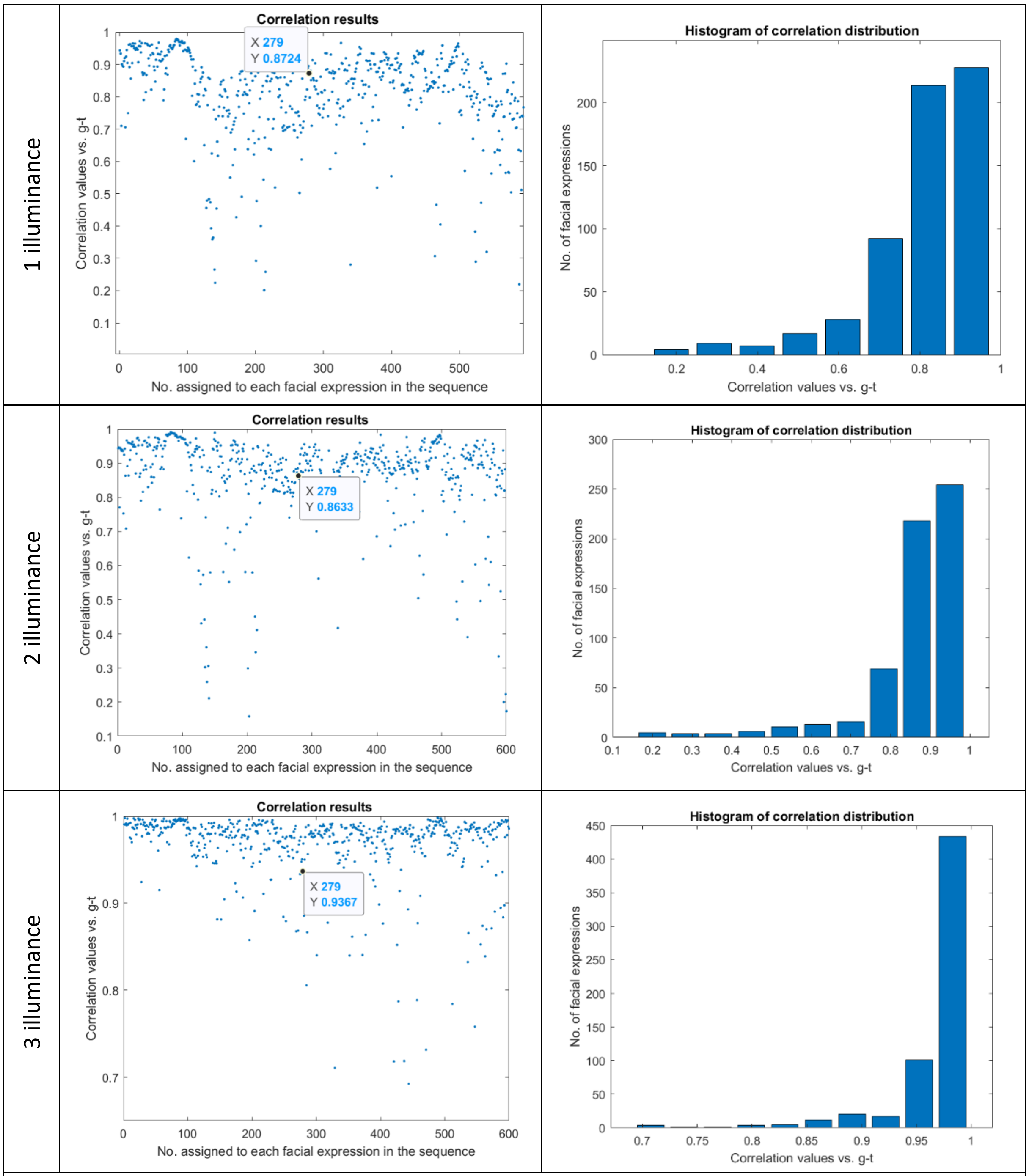
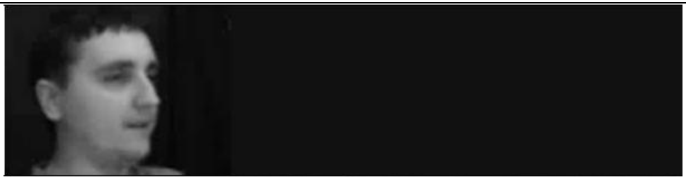
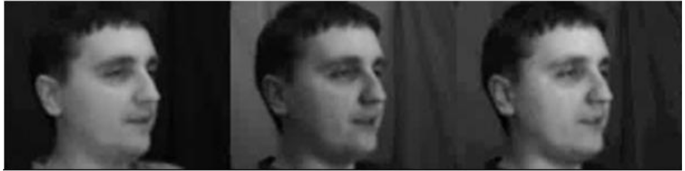
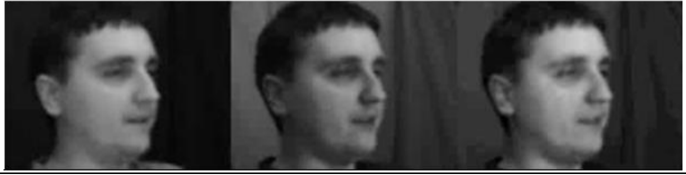


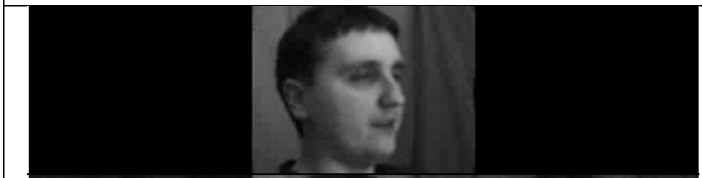
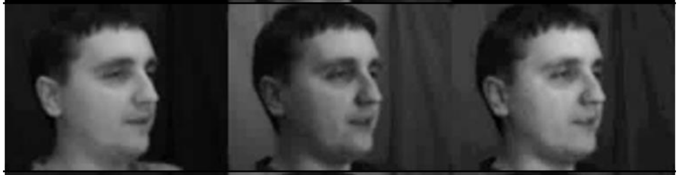
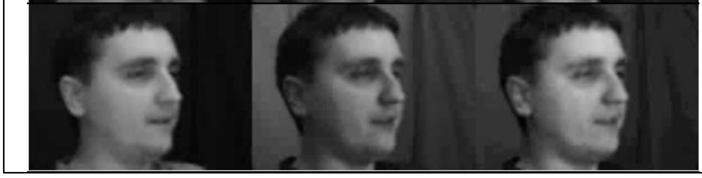
Figure 77 shows the correlation results. The points in three graphs of the left column show PC scores relation that defines how close the scaled/reconstructed 1st illuminant, 2nd illuminant, and 3rd illuminant are to the ground-truth. Also, a point in these graphs, at position 279 shows the correlation between the scaled/reconstructed 1st, 2nd and 3rd illuminants and the ground-truth accordingly. The blue bars in three graphs of the right column show the distribution of correlation between pairs of image weights of reconstructed illuminants and the ground-truth data**.**

5.2.4. Fourth illuminance condition

For the fourth experiment, we applied a controlled uniform light and repeated the steps from the first experiment. The results of the second experiment showed that the missing illuminated faces could be correctly reconstructed, by the system, from only one illuminated face, as can be observed in **Figure 78**.

Uniform lighting

1st Illuminant	
	The 1st illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 1st illuminant data image information.
	Ground-truth multi-illuminant image representation.

2nd Illuminant	
	The 2nd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 2nd illuminant data image information.
	Ground-truth multi-illuminant image representation.

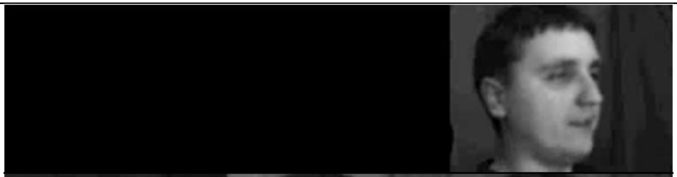
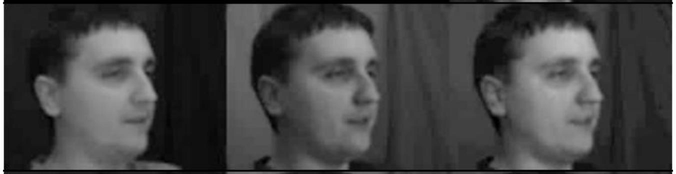
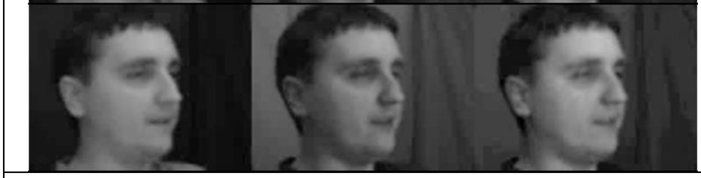
3rd Illuminant	
	The 3rd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 3rd illuminant data image information.
	Ground-truth multi-illuminant image representation.

Figure 78 *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* show reconstruction of different facial expressions having only information of the face from *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* (see **Movie 10**.avi).

The multi-illuminant images in **Figure 78** are the results of the face reconstruction from a single illuminant. From visual inspection of **Figure 78,** reconstructions from all illuminants appear identical to the ground-truth. In **Figure 79,** we may observe on the graph that our images are similar to the ground-truth data. The reconstructed data from the 3rd illuminant is quite close to the ground-truth. Meanwhile, the reconstructed data from the 1st and 2nd illuminant are slightly muted but show a better match than in previous experiments, and the other two provide a very weak match with ground-truth. However, we will repeat the procedure to find a scaling factor that will give the best match with ground-truth images.

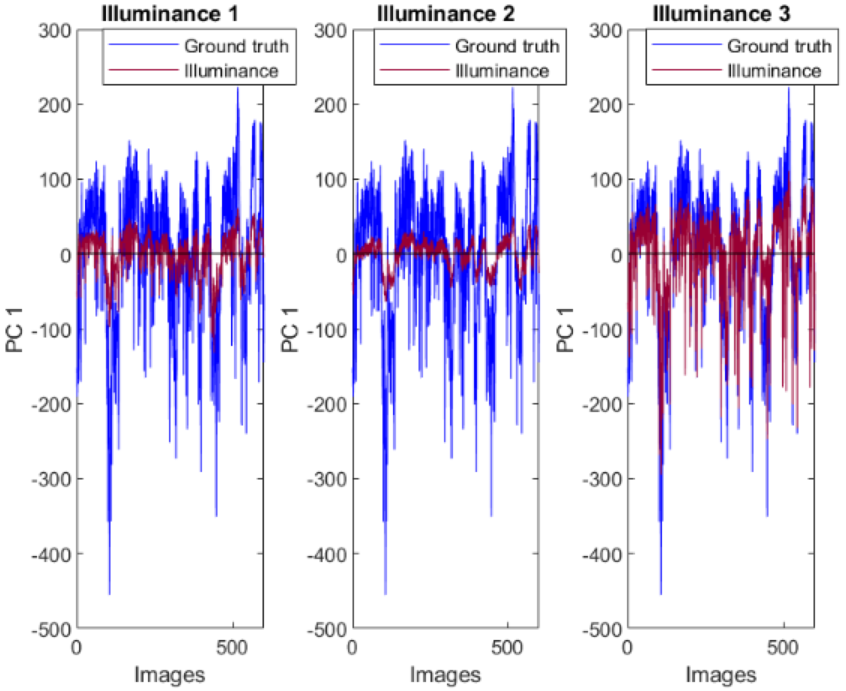


Figure 79 all three graphs represent loadings comparison of reconstructed multi-illuminant images with ground-truth representation. The graph on the left shows how loadings of reconstructed multi-illuminant images from 1st illuminant (red colour) are close to the ground-truth (blue colour) multi-illuminant images. The middle graph shows a loadings distribution for reconstructed multi-illuminant from 2nd illuminant (red colour) and ground-truth (blue colour). The right graph present loadings distribution of reconstructed multi-illuminant images from 3rd illuminant (red colour) with ground-truth multi-illuminant images (blue colour). As we may notice from the graphs, the loadings of reconstructed multi-illuminant images from the 3rd illuminant are very close to the ground-truth multi-illuminant representation; it means that images should be very similar to each other. The loadings of the other two graphs are less close to the ground-truth, hence reconstructed images should contain fewer similarities with ground-truth images.

From **Figure 80,** we may see that a scaling factor for the 3rd illuminant reconstructed view range is smaller than a scaling factor range for the 1st and 2nd illuminant. That is because the 1st and 2nd illuminant are a bit darker than the 3rd illuminant image. Hence their scaling factor will have to be bigger to match ground-truth.

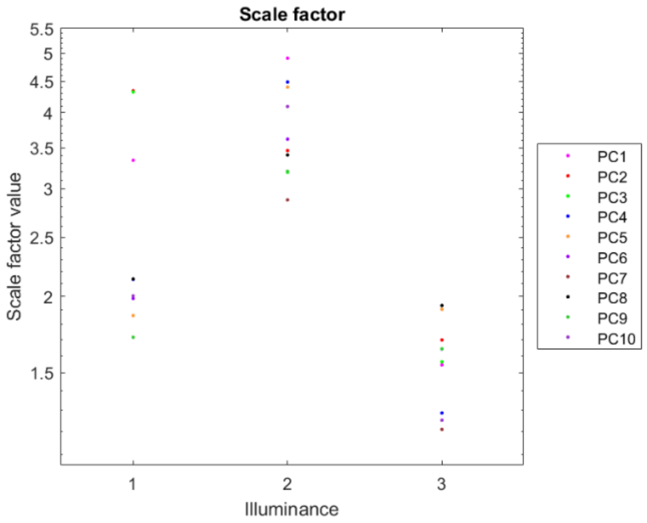


Figure 80 The figure presents a scaling factor data distribution for the 1st, 2nd and 3rd illuminant. As we may see, the largest data range is for 1st illuminant. The 2nd illuminant scaling factor data range is smaller than the 1st illuminant. The 3rd illuminant has the smallest scaling factor.

Figure 80 shows that the scaling factor for the 3rd illuminant reconstructed view is much smaller than the scaling factor for the 1st and 2nd illuminant. That is because the 1st and 2nd illuminant are slightly darker than the (bright) 3rd illuminant, hence their scaling factor will need to be bigger to match ground-truth.

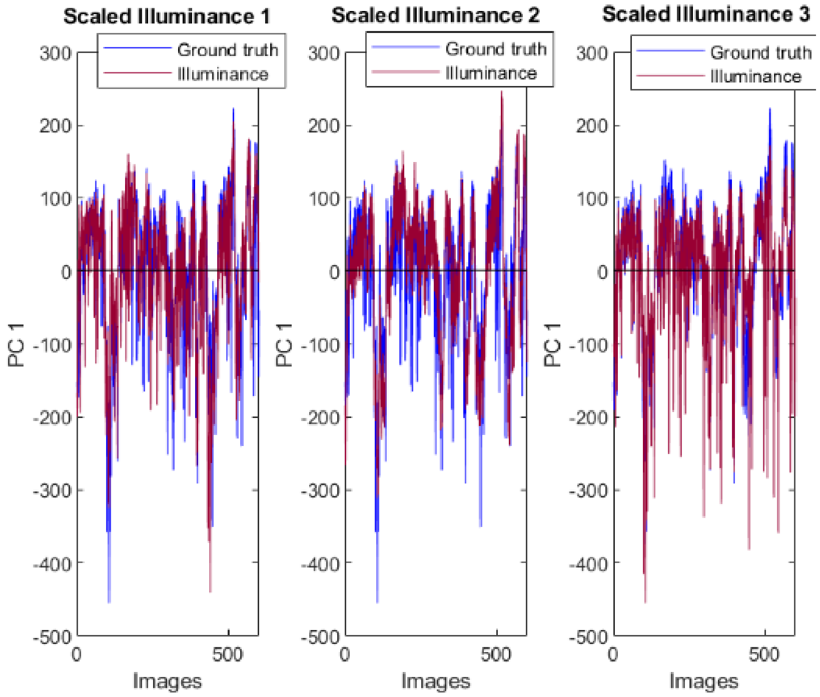
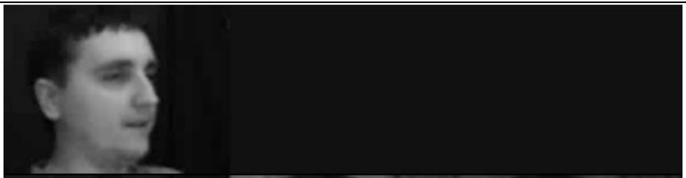
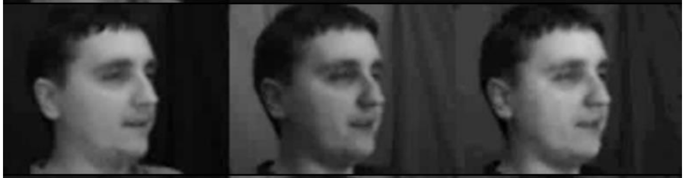
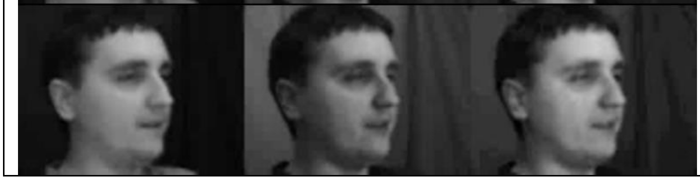
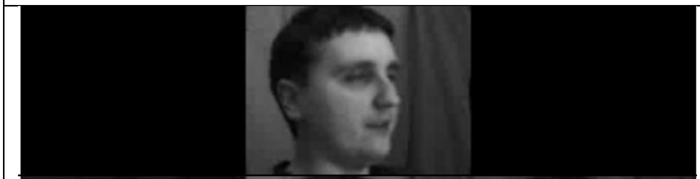
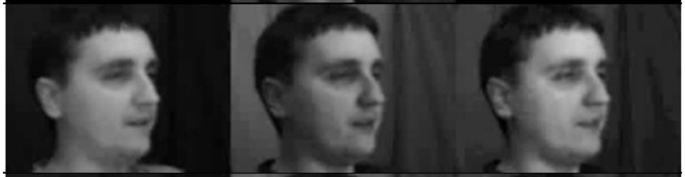
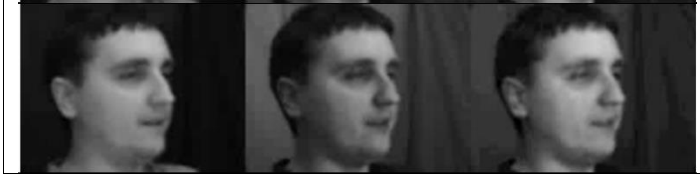


Figure 81 The graph presents loadings matching of scaled loadings from 1st, 2nd and 3rd illuminant (red colour) with ground-truth loadings (blue colour). As we may notice from the graphs, scaled loadings gave a better match to ground-truth.

In **Figure 81,** we may observe scaled loadings that are similar to the ground-truth. The results of reconstructed images after applying the scaling factors are presented in **Figure 82**.

Uniform lighting

1st Illuminant	
	The 1st illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 1st illuminant data then applying a scaling factor to match a ground-truth representation.
	Ground-truth multi-illuminant image representation.

2nd Illuminant	
	The 2nd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 2nd illuminant data then applying a scaling factor to match a ground-truth representation.
	Ground-truth multi-illuminant image representation.

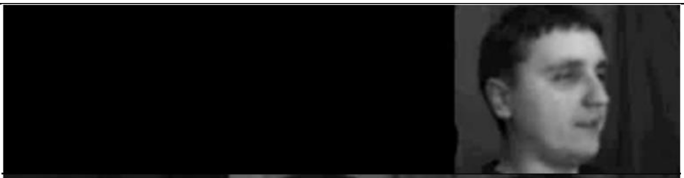
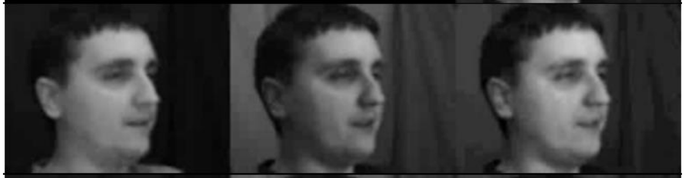
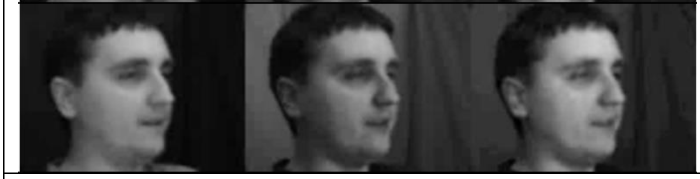
3rd Illuminant	
	The 3rd illuminant data in the multi-illuminant image is present. The rest information is taken away (blacked out).
	The multi-illuminant image reconstructed using only 3rd illuminant data then applying a scaling factor to match a ground-truth representation.
	Ground-truth multi-illuminant image representation.

Figure 82 *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* show reconstruction of different facial expressions having only information of the face from *1st Illuminant*, *2nd Illuminant* and *3rd Illuminant* (see **Movie 11**.avi).

From visual inspection in **Figure 82,** it can be noticed that the reconstructed image for all illuminants after applying a scaling factor looks very close to the ground-truth representation from our content addressable memory. However, to the naked eye, the reconstructed images looked the same as the ground-truth after applying a scaling factor. The only way we could see the difference between the images was by checking the data from the graphs. The accuracy was to be expected because the scaling factors have been obtained individually for each principal component and illuminant.

By examining the correlation coefficients of facial expression at position 281 in graphs of the left column in **Figure 83,** we may see a very high correlation value between scaled/reconstructed data and ground-truth. As in the first, second and third experiments, the correlation exceeds 0.8. This means that the reconstructed and later scaled images derived from multiple illuminants are a close match to the original data representation. In the histogram graphs of the right column, we may observe how correlation data is distributed for every illuminant.

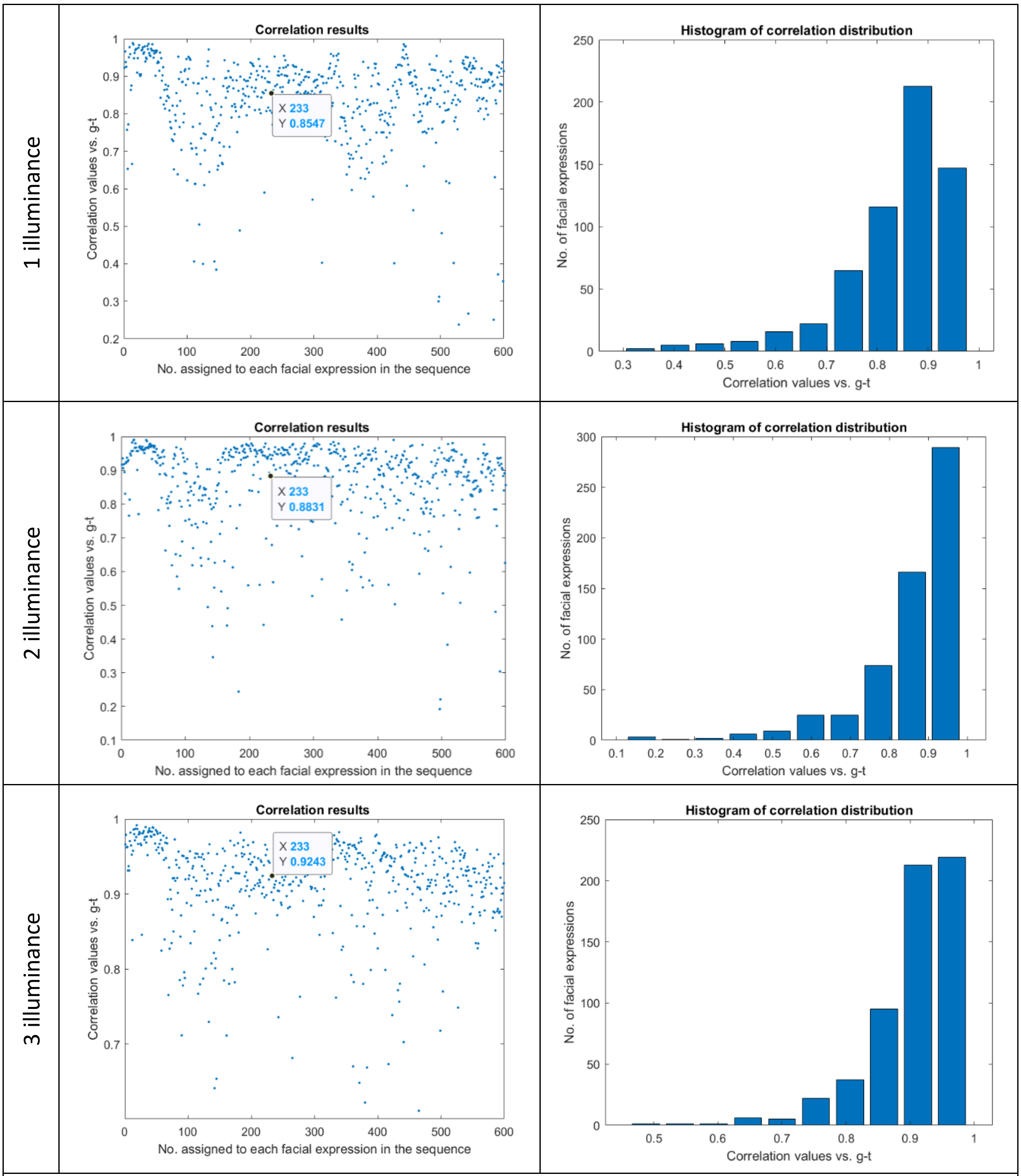


Figure 83 shows correlation results. The points in three graphs of the left column show PC scores relation that defines how close the scaled/reconstructed 1st illuminant, 2nd illuminant, and 3rd illuminant are to the ground-truth. Also, a point in these graphs, at position 233 shows the correlation between the scaled/reconstructed 1st, 2nd and 3rd illuminants and the ground-truth accordingly. The blue bars in three graphs of the right column show the distribution of correlation between pairs of image weights of reconstructed illuminants and the ground-truth data.

Our experiment in four illuminance conditions demonstrates that we can extend the set of existing techniques for generating a representation that allows for the encoding of multiple views in a single representation to the problem of lighting invariance. We know that changes in illumination can cause a dramatic variation in the appearance of the facial image and seriously affect the performance of face recognition systems. The standard way to achieve illumination invariance in computer vision is to eliminate the influence of the illumination by operating on edges, thereby removing any shading of the face. Acknowledging the fact that humans can recognise faces in various illuminants when only a part of the face is present or when the face has been binarized, e.g. Mooney faces, we have addressed the problem of illuminant invariance for grey-scale facial images. In this work, we have included all the shading information in the face and analysed the face holistically. We found that this general approach could relate faces captured under different illumination conditions. To test the model, we created four non-rigid facial sequences with different types of diffuse lighting positions. Every facial motion sequence was captured as a multi-illuminant image, consisting of three different illuminants. The diffuse lights from the left and right had some effect on the reconstruction of the data. However, the reconstructions still have better correlations than one would get by looking at random pairings of the ground-truth data. The reconstructions had defects that were visible to the naked eye. From visual inspection, the reconstructed multi-illuminant images from the uniform lighting condition before applying a scaling factor looked identical to the original data representation.

Results showed that the best match between reconstructed and ground-truth data was obtained from the 3rd illuminant. When the illuminant was weak, the reconstruction results were weaker, e.g. data derived from the 1st and 2nd illuminants. Nevertheless, when we amplified the muted reconstruction of facial behaviour for different illuminants by scaling against ground-truth, the reconstructions were successful. We found correlation coefficients that define how well the reconstructed results are in relation to the original data. The correlation coefficients of four controlled lighting situations for all illuminants of randomly chosen face expression from the sequence ranged between 0.8 and 0.9. It means that the reconstructed images are almost identical to the original data. This method helped to obtain an answer to the problem of recognising/reconstructing face over different lighting conditions.

5.3. Chapter 5 summary

In chapter 4, a face recognition problem of multiple poses has been introduced, which was solved using a view-based approach built on the existing Cowe PCA-based model (Cowe 2003). In chapter 5, we wanted to go further and apply a view-based approach to address the illuminance problem. Here hardware has been modified (see in 3.1.3) so that we would achieve the illusion of different lights in different locations in the face. The main reason why we have chosen to create and implement this method to solve the lighting invariance problem is that we can build a set up for every single frame and receive three different illuminances per every single frame. There is no switching between illuminances on a frame by frame basis, which would be very complicated to do.

6. Object constancy over asynchronous view-based data using PCA-based facial mimicry.

Humans have an excellent capability to distinguish between faces. However, they are good at recognising or matching faces that are familiar to them, even if they are presented in very low-quality images (Burton et al., 1999; Natu & O'Toole, 2011), but are much worse at recognising, or merely matching unfamiliar faces (Bruce et al., 2001; Burton et al., 1999; Megreya & Burton, 2006). The process of learning a new face requires perceptive flexibility of forming that person's face in different representations. In previous chapters (chapter 4 and chapter 5), we have investigated face recognition in multiple pose and lighting conditions and were able to reference multiple appearances of the same face to a single general cause using a multi-vector with the use of the PCA-based approach. However, humans cannot see all the different face views at the same time, and neither can they observe the variety of lighting conditions at the same time. Therefore, the principal aim of this chapter is to explore whether it is possible to build a multi-view model vector from asynchronous data to handle unfamiliar face reconstruction at wide angular differences between viewpoints. This would help us to achieve object constancy for unfamiliar faces across multiple views.

Multiple research showed that change in viewpoint has a significant influence on the recognition of unfamiliar faces (e.g., Bruce, 1982; O'Toole et al., 1998). Bruce, in his research, discovered that when the views of faces between study and test phases were different, recognition time and accuracy for unfamiliar faces was lower than for familiar faces (Bruce, 1982). He noted that recognition decreased from 90 per cent when views of the faces in both phases were the same to 76 per cent and when they were different. In addition, participants in this study had a better recognition memory for unfamiliar faces when they had prior knowledge of face pose while performing face matching tasks other studies showed that familiar faces could be recognised even under significant changes in pose (Bruce & Young, 1986; Eger et al., 2005; Hill et al., 1997). These findings suggest that perception and recognition of unfamiliar faces are viewpoint dependent when familiar faces appear to be represented in a viewpoint-invariant manner. Several studies confirmed that by revealing that familiar and, unfamiliar faces have different neural mechanisms regarding the viewpoint information (Pourtois et al., 2005; Ryu & Chaudhuri, 2006). Pourtois et al. (2005), in their study, observed a repetition-related activation decrease on the right side of the fusiform cortex for unfamiliar faces. For familiar faces, the same specific region was found in the left medial of the fusiform gyrus. This leads to a conclusion that although recognition of familiar faces is viewpoint-invariant, the processing of familiar faces still depends on viewpoint information.

Other studies claim that changes in facial expressions could have a significantly higher impact on the recognition of unfamiliar faces than of familiar faces. Bruce (1982) used different viewpoints and expressions for the study and test phases and found that the resulting performance in recognition of unfamiliar faces reached only 61 per cent (Bruce, 1982). The reaction times in matching faces with different facial expressions, on the other hand, were the same for familiar and unfamiliar faces. Nevertheless, if one eliminates the information of facial expressions from the face-matching process, the effect of familiarity can still be observed. Although the facial expressions did not affect the reaction time of the participants while matching unfamiliar faces, the reaction time for matching familiar faces was much faster with faces that represented happy expressions than with faces that represented angry expressions (Kaufmann & Schweinberger, 2004). It shows that even recognition of familiar faces depends on some expression information to aid the recognition process.

Surrounding information (in rigid changes in viewing angle) also impacts face perception. However, it has a larger effect on the perception of an unfamiliar face than a familiar one (Memon & Bruce, 1985). The recognition-memory task showed that by changing both face pose and surrounding information, one could significantly impair perception of unfamiliar faces, while the perception of familiar faces will stay unaffected (Davies & Milne, 1982). Other studies also observed the negative influence of surrounding information on the recognition of unfamiliar faces (Dalton, 1993; Russo et al., 1999).

As we see, image changes such as variation in the angle of viewing, facial expressions, or external information are very harmful to unfamiliar face matching and recognition performance tasks. On the other hand, recognition performance stays very high for familiar faces regardless of such changes. Seeing faces in various viewing angles or facial expressions is essential in building a representation of familiar faces. Familiar faces are represented by internal features of a face, which also helps effective face perception. Unfortunately, unfamiliar faces are viewpoint dependent and cannot rely on internal features information in a way as familiar faces can. It seems that variations in the images extremely influence the recognition of unfamiliar faces. Thus, it is very important to have extreme caution and clarity while choosing information that will be used in the images in the comparison task of unfamiliar faces. Several studies confirmed such sensitivity to the changes in the images. For example, Burton et al. showed that participants could easily recognise and match familiar faces even in low-quality commercial CCTV images and failed with matching unfamiliar familiar faces (Bruce et al., 2001; Burton et al., 1999). Bruce et al. showed that participants performed poorly in the face determination task even when they had prior knowledge of an unfamiliar face and were hesitant to answer whether the face they have seen before could be present in one of the ten images that were laid in front of them. The recognition accuracy for this task reached only 70 per cent. Even when participants were told that a person they have previously seen was present on the sample images, the accuracy was still only 79 per cent.

In this chapter, we examine whether it is possible to build and use a multi-view model vector from asynchronous data to handle unfamiliar face reconstruction at vast angular differences between viewpoints and include all of the above-mentioned variations that disturb the matching performance of unfamiliar faces. In order to address this issue, PCA spaces will be constructed separately for a frontal face, three quarter, two third and profile views. We will then take a face in a frontal view with a particular facial expression and generate this face in the other PCA spaces. Next, we will group the data together to make a surrogate multi-vector, which would be almost equivalent to recording the different face views at the same time. Last, we will compare the data with the ground-truth to see how well this strategy worked.

6.1. Test stimuli: the database of asynchronous face sequences.

A set of non-rigid facial motion sequences was created to test the model. The sequences represent a non-rigid capture of facial behaviour from four different viewpoints filmed under the same lighting conditions at the same times. Sequences were captured using four Grasshopper GRAS-03K2C (FireWire) digital cameras, at a rate of 60 frames per second, at 640x480 image resolution with RGB 24-bit pixel format. The sequences were of an expressive talking face, and they lasted around 30 seconds. For this experiment, the asynchronous data sequence was constructed from synchronised data sequences. The motion sequences were aligned, as described in 3.1., and calibrated, as described in 3.1.1., and 3.1.2. The sequences were vectorised as described in 3.3.1. The sequences for each face pose were processed with PCA, and the basis vectors were extracted to form a PCA-based model. The computational strategy that was used to create the asynchronous data is introduced below.

6.2. Results of the experiment

The results of the first experiment show the reconstruction of four perspective views of a face from a single perspective. For this experiment, four separate PCA spaces were created, one for each unique viewpoint of the face (e.g. frontal view, 3/4 view, 2/3 view and profile view). A frontal view of the face was taken as an example and was mapped into these different PCA spaces separately to find an image that would best represent the data in the different poses from the input face.

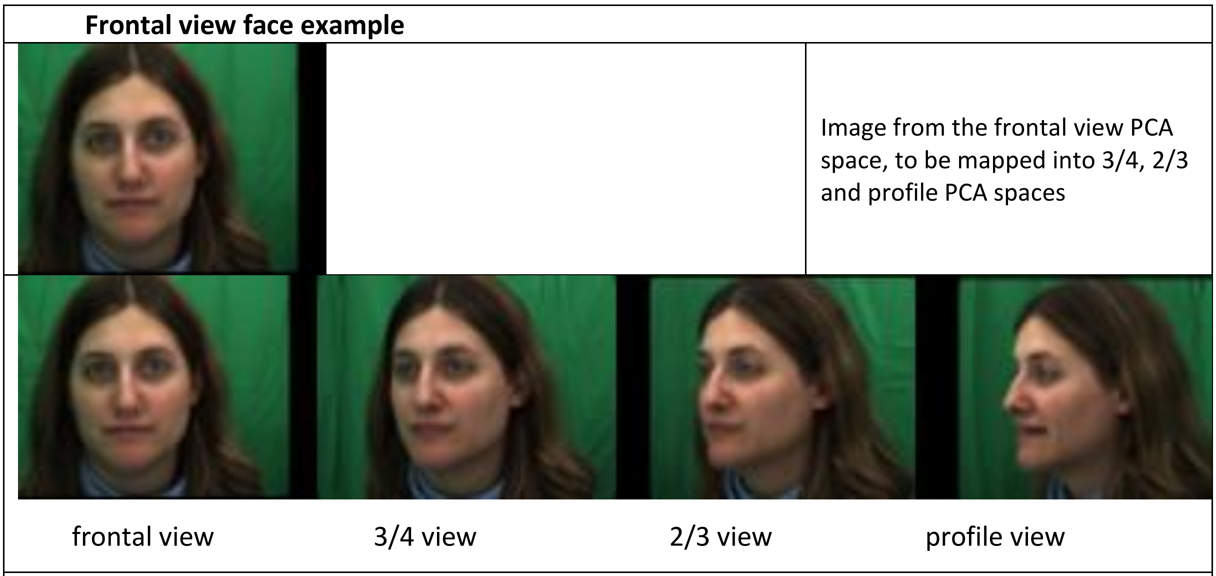
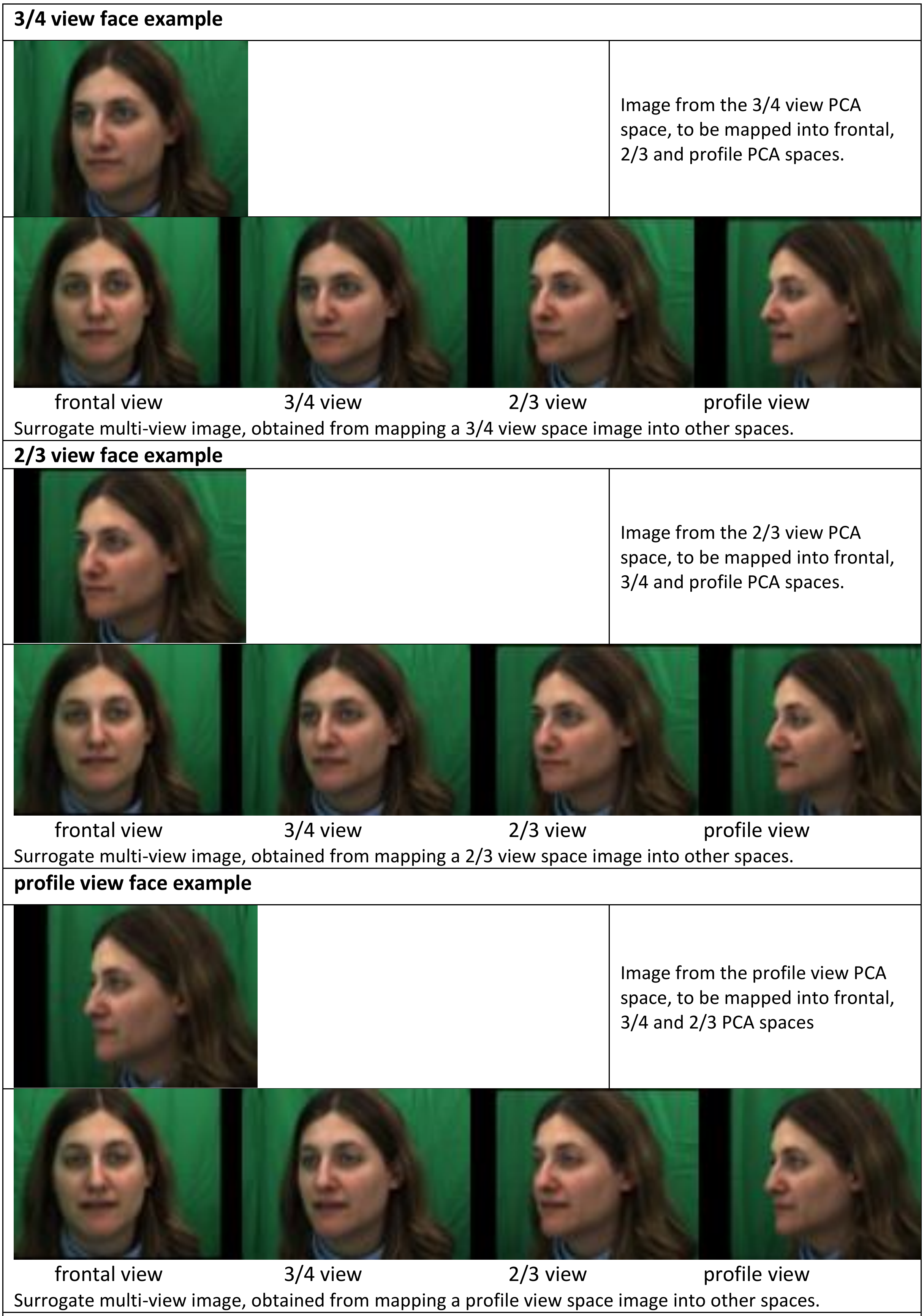


Image from the frontal view PCA space, to be mapped into 3/4, 2/3 and profile PCA spaces

Figure 84 Face expression of frontal view reconstructed in other perspectives. Reconstructed looks were concatenated together to build one multi-view surrogate image. A surrogate multi-view image was obtained from separate mapping frontal view space image into 3/4 view, 2/3 view, and profile view PCA spaces.

Figure 84 demonstrates an asynchronous surrogate multi-view image created from the mapping of a frontal view PCA space image into 3/4, 2/3, and profile PCA spaces. As mentioned earlier, synchronised data was used to create asynchronous surrogate multi-view representation; however, a PCA space for every viewpoint was created from synchronised data sequences collected at different time points. Hence, views in the multi-view image are not related. The frontal view is 18 degrees away from 3/4 view, 36 degrees away from 2/3 view and 54 degrees away from a profile view. Here, the asynchronous multi-view images can only be judged by eye because there is no ground-truth representation of the faces with the mapped expressions. From visual inspection, the frontal face expression of eyes wide open and mouth shut appear to be in all the mapped views. However, if we take a closer look at the facial expression of the profile view, the mouth is more open, but the eyes are almost closed. It was an unexpected mapping behaviour. The 3/4 face multi-view appear to be similar to the frontal face multi-view, with a small error in the frontal face, where eyes appear to be slightly closed. The 2/3 view gave slightly worse results than the 3/4 view in the reconstruction of facial expression. The eyes and the mouth in the 2/3 view were more closed than on the face in the 3/4 view. How would surrogate multi-view face image look like if mapping would be made from 3/4, 2/3, and profile views? (see **Figure 85**)



3/4 view face example

Image from the 3/4 view PCA space, to be mapped into frontal, 2/3 and profile PCA spaces.

Surrogate multi-view image, obtained from mapping a 3/4 view space image into other spaces.

2/3 view face example

Image from the 2/3 view PCA space, to be mapped into frontal, 3/4 and profile PCA spaces.

Surrogate multi-view image, obtained from mapping a 2/3 view space image into other spaces.

profile view face example

Image from the profile view PCA space, to be mapped into frontal, 3/4 and 2/3 PCA spaces

Surrogate multi-view image, obtained from mapping a profile view space image into other spaces.

Figure 85 Face expression of one view reconstructed in other views.

Figure 85 presents the surrogate multi-view face images made from mapping 3/4, 2/3 and profile view spaces into the other view spaces. From visual inspection, the 3/4 face multi-view appears to be similar to the frontal face multi-view, with a small error in the frontal face, where eyes appear to be slightly closed. The 2/3 face multi-view gave slightly better results in the reconstruction of the facial expression than the profile view. The mouth is slightly opened, and the eyes are somewhat shut in all of the reconstructions from 2/3 face multi-view image. The profile view had little information about how the facial expression should look like from other perspectives; hence, nearly closed eyes and open mouth can be observed in mapped facial expressions of all views.

Having built separate PCA spaces, we wanted to go further and test the multi-view model with asynchronous data providing object constancy over multiple views in a more natural setting. Hence, the surrogate multi-view asynchronous image sequence was vectorised as described in 3.3.1. The sequence was processed using PCA, and the basis vectors were extracted to form a PCA-based model. Next, a reconstruction method was used, which was introduced to investigate object constancy over multiple poses and in different illuminance conditions (Experiment 3 in chapter 4 and Experiment 1 in chapter 5). The content addressable memory was loaded with 446 asynchronous multi-views, which included a wide range of views and non-rigid facial expressions. The content addressable memory (PCA space with all views) was used to reconstruct missing viewpoints in a multi-view image (see **Figure 86**).

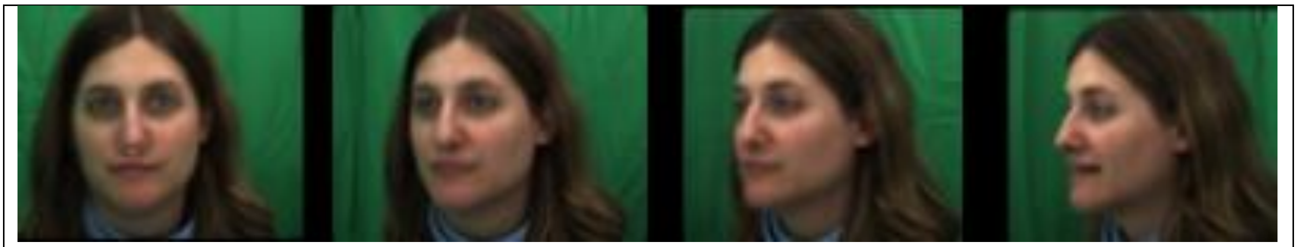
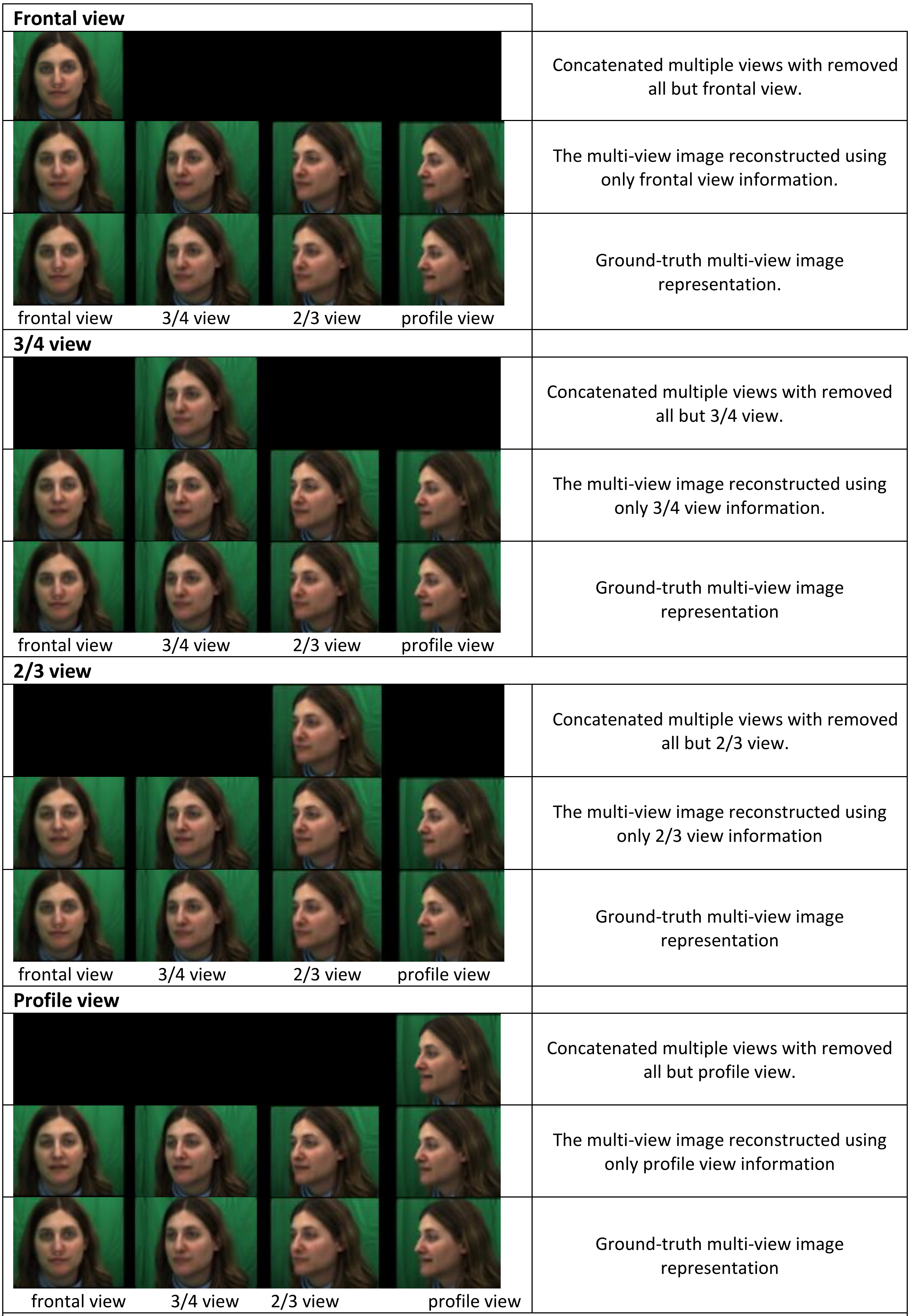


Figure 86 One of 466 concatenated multiple views, which are stored in the content addressable memory.

After creating a PCA space from the asynchronous sequence, the nose area became blurred in the frontal face representation. To reconstruct missing asynchronous face views from only one face view, we mapped an image vector derived from that face into a multi-view content addressable memory. The reconstruction of missing faces was performed from frontal, 3/4, 2/3 and profile views. The process of reconstructing missing asynchronous face views was repeated as introduced in subsection 4.1.5 and chapter 5. The reconstruction results are presented in **Figure 87**.



Frontal view

Concatenated multiple views with removed all but frontal view.

The multi-view image reconstructed using only frontal view information.

Ground-truth multi-view image representation.

3/4 view

Concatenated multiple views with removed all but 3/4 view.

The multi-view image reconstructed using only 3/4 view information.

Ground-truth multi-view image representation

2/3 view

Concatenated multiple views with removed all but 2/3 view.

The multi-view image reconstructed using only 2/3 view information

Ground-truth multi-view image representation

Profile view

Concatenated multiple views with removed all but profile view.

The multi-view image reconstructed using only profile view information

Ground-truth multi-view image representation

Figure 87 Face reconstruction over multiple views using only certain viewpoint information (see **Movie 12**.avi).

Rigid changes in the face were better reconstructed from all views than non-rigid facial information. Reconstructed views in multi-view image from the frontal view looks very similar to an example perspective (see **Figure 87**). The blur in the nose area disappeared for the frontal view. The open eyes and a shut mouth appear to be almost in all reconstructed views, except profile view, where the mouth is slightly open. From a closer look, starting from the frontal view and moving to the profile view, the eyes of the person are gradually beginning to close. Similar results are noted for reconstructed multi-view image from 3/4, 2/3 and profile perspectives. A similar tendency was noted in data reconstruction in Experiment 3 of chapter 4 and Experiment 1,2,3 of chapter 5. To summarise, the results of the visual inspection show that the best match with ground-truth was obtained from the frontal view. However, it is hard to identify which viewpoint gave the closest match to the ground-truth with visual inspection only. Hence, we have plotted the 1st principal component values for both the reconstructed and the ground-truth face representations. It was done to have a better observation of how close reconstructed images are to the ground-truth (see **Figure 88**).

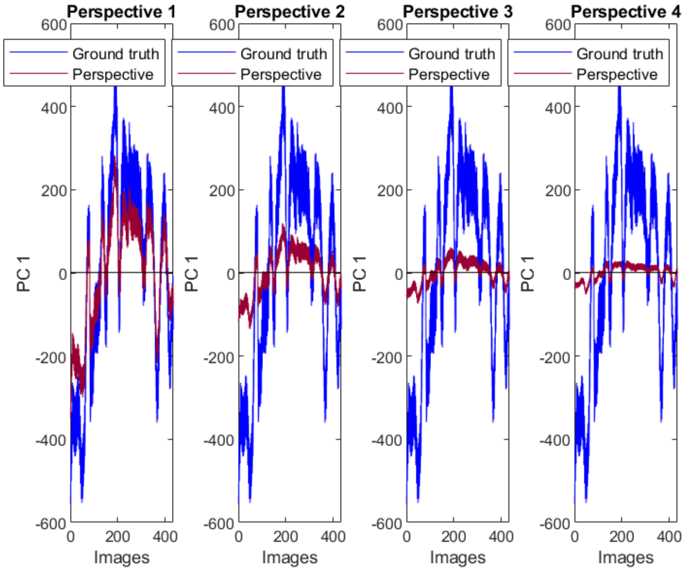


Figure 88 Reconstructed and ground-truth of the 1st principal component values.

From **Figure 88,** we may see that the best face reconstruction with ground-truth was made from the frontal view (left graph). The worst match with the ground-truth was made from the profile view. We have chosen to enhance the reconstructed values to match our ground-truth representation. A linear least-squares method was used to find the scale factor that generates the best correspondence with the ground-truth loadings. The obtained scaling factor for every principal component (ten principal components) and viewpoint is presented in **Figure 89.**

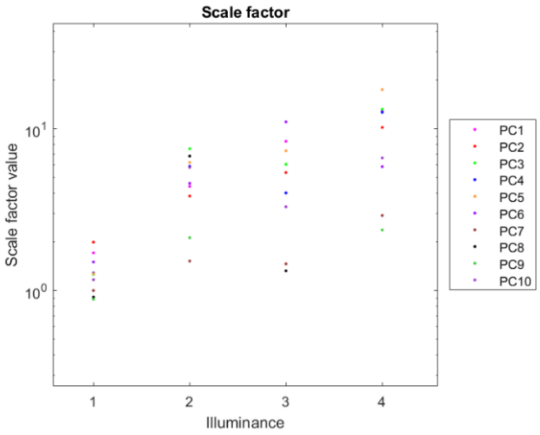


Figure 89 The scaling factor for four views to match the original data representation.

From **Figure 89,** we observe that for the frontal view (**number 1**) the scaling factor is much smaller than for the rest of the views. The scaling factors for the profile (**number 4**) and 2/3 view (**number 3**) are scattered the most because these two views contain less dynamic facial information than the frontal (**number 1**) and 3/4 (**number 2**) views. Next, the obtained scaling factor was applied to the reconstructed data.

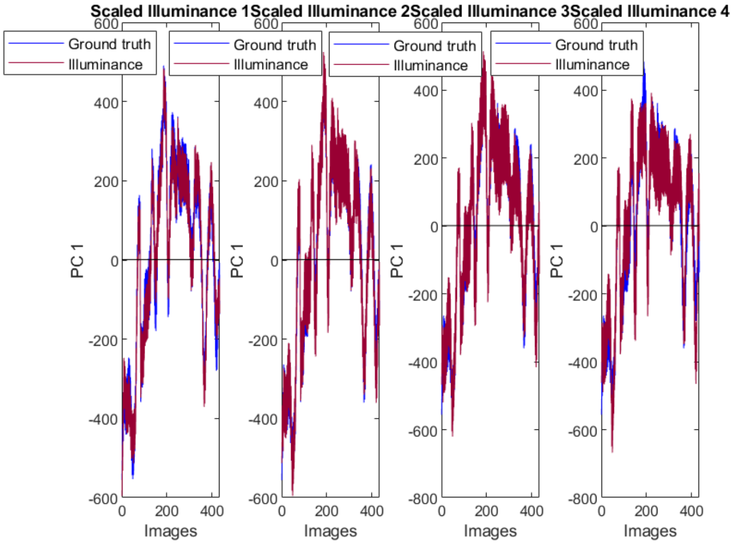


Figure 90 Reconstructed and ground-truth of the 1st (plot on the left) and 5th (plot on the right)principal component values are after applying a scaling factor.

In **Figure 90,** we can compare the reconstructed and ground-truth values of the 1st PC component after applying a scaling factor. We can see that all reconstructed multi-views (red line) over all graphs after scaling had a better match with the ground-truth representation (blue line). However, we can also observe that in **Figure 90,** the profile and 2/3 views provided less than a perfect match to the ground-truth after the scaling factor was applied. In the profile view graph (Scaled perspective 4), the reconstructed data do not reach the ground-truth representation. For the 2/3 view (Scaled perspective 3), we may see some data exaggeration. For the rest of the graphs, the scaled data has a better match with the ground-truth. We can also evaluate these results at the image level (**Figure 91**).

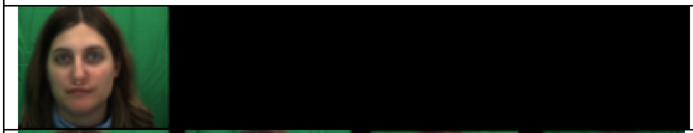
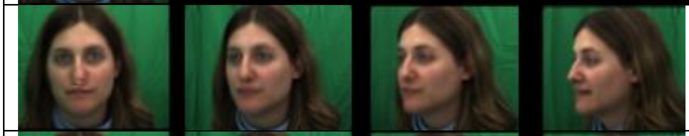
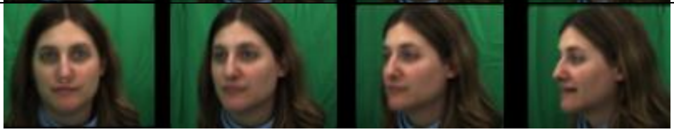
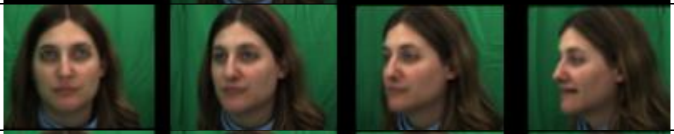
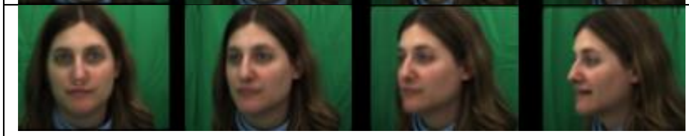
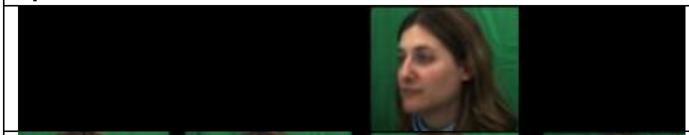
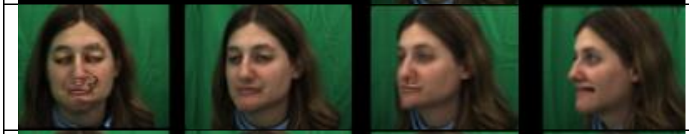
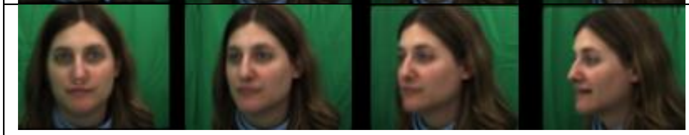
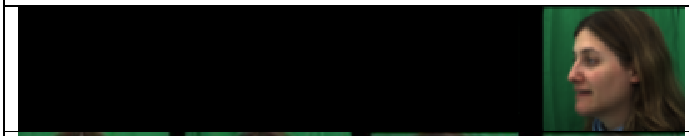
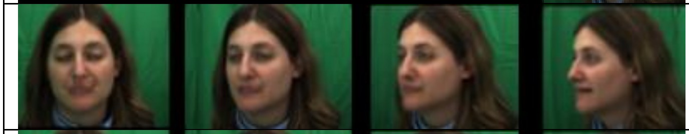
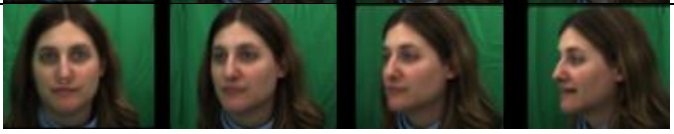
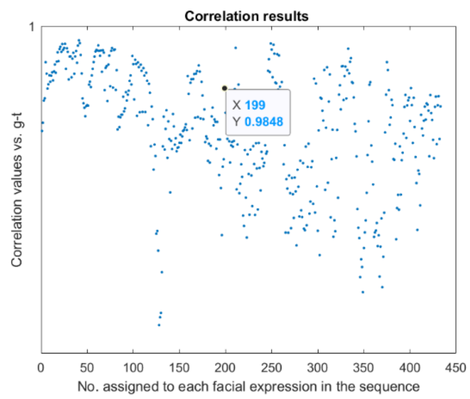
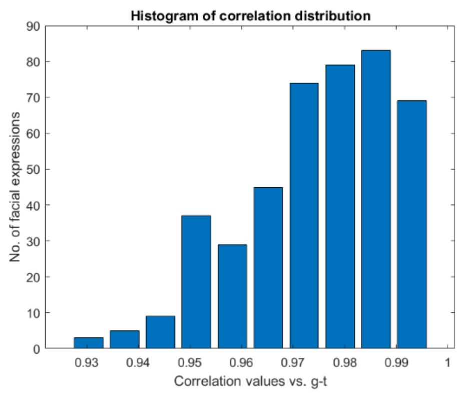
Frontal view	
	Concatenated multiple views with removed all but frontal view.
	The multi-view image reconstructed using only frontal view information.
	Ground-truth multi-view image representation
3/4 view	
	Concatenated multiple views with removed all but 3/4 view.
	The multi-view image reconstructed using only 3/4 view information.
	Ground-truth multi-view image representation
2/3 view	
	Concatenated multiple views with removed all but 2/3 view.
	The multi-view image reconstructed using only 2/3 view information
	Ground-truth multi-view image representation
Profile view	
	Concatenated multiple views with removed all but profile view.
	The multi-view image reconstructed using only profile view information
	Ground-truth multi-view image representation

Figure 91 Face reconstruction over multiple views using only one specified viewpoint after applying a scaling factor to match original data. (see **Movie 13**.avi).

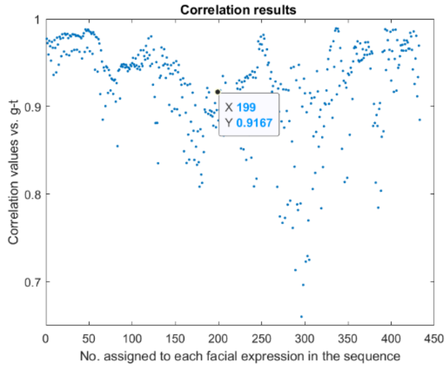
From visual inspection, it can be noticed that the facial reconstruction for the frontal view after applying a scaling factor looks accurate and is very close to the ground-truth representation stored in the PCA content addressable memory. However, it can also be noticed that the face reconstructions of the other views look poor compared to the ground-truth. In the 3/4 view multi-view results, we can see an exaggeration of the opening of the eyes on the frontal face. In the 2/3 view reconstruction, we can observe some image disturbance for all the views. The profile view appears to be muted and imprecise in comparison to the ground-truth representation. However, from visual inspection, it appears that the reconstruction results obtained from the profile view were better than those from the 2/3 view, which was unexpected. Therefore, to better evaluate obtained results, we measured how all reconstructed and scaled multi-view images related to the ground-truth using the standard Pearson product-moment correlation coefficient approach. The mimicry generation and the degree of correlation between the reconstructed face pose and the ground-truth representation are presented in **Figure 92**. We found a correlation between principal component loadings extracted from the 433 images that were black out at a specific part of multi-view driver sequence vectors and those from the ground-truth. The correlation values for frontal view were ranging between 0.927 to 0.9965, for the 3/4 view – 0.6599 to 0.9894, for the 2/3 view – 0.2 to 0.97 and for the profile view – 0.2 to 0.9756. We have measured the percentage of data that had correlation values above 0.8. The data from frontal view had 100 per cent, 3/4 view had 99,7 per cent, 2/3 view had 9 per cent, and profile view had 14 per cent of data with correlation coefficients above 0.8. These results suggest that the frontal face view have the most facial information to achieve the face reconstruction closest to the ground-truth, whereas the 3/4 view is next best.

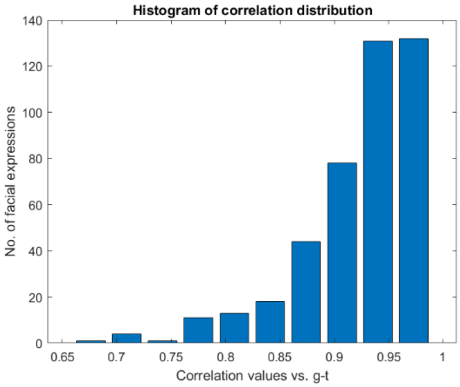
Interestingly, the profile view performed better in the face reconstruction task than the 2/3 view. This is an anomaly that was unexpected, and it will be discussed later in this chapter. Next, we have randomly chosen an image from the sequence at position 199 to evaluate correlation coefficients between the reconstructed multi-view image from frontal, 3/4, 2/3, profile views and ground-truth accordingly (see **Figure 92**, graphs on the left column). The correlation values were ranging from 0.6648 to 0.9848. From graphs on the left column in **Figure 92,** we can see that the highest reconstruction value was obtained from the frontal view with the value of 0.9848, the second largest value was obtained from the 3/4 view - 0.9167. The profile view gave the third most value, 0.6866, and the least value was obtained from the 2/3 perspective – 0.6648 frontal view



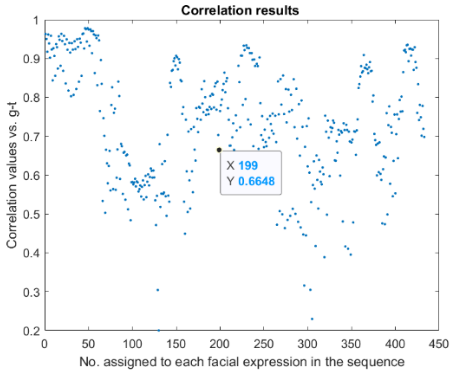


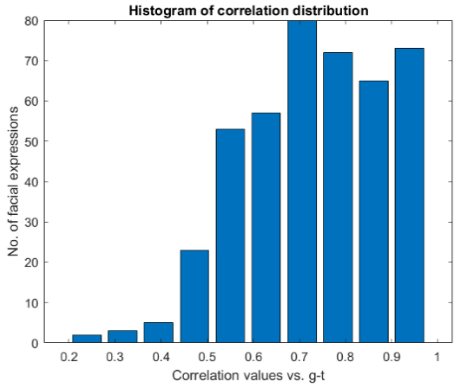
3/4 view



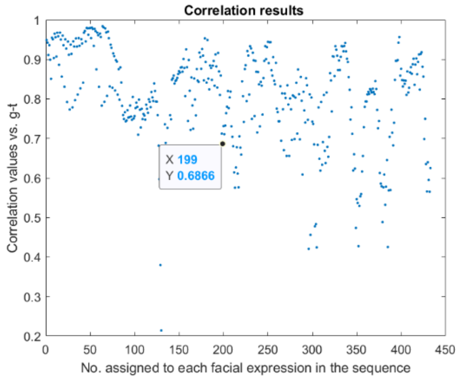


2/3 view





profile view



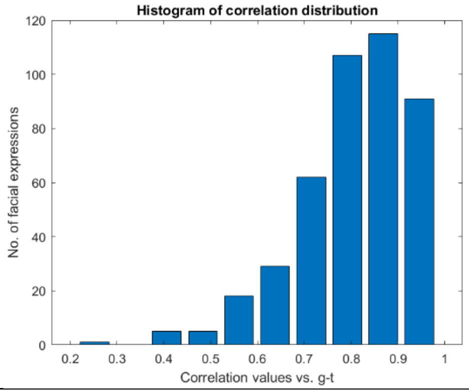


Figure 92 Correlation results for the multi-view face reconstruction using only frontal, 3/4, 2/3 and profile views information accordingly. The points in plots on the left column denotes loadings that show how close the reconstruction results are to the original data. Also, a point in the plots, at position 199, presents the correlation between the reconstructed multi-view image from frontal, 3/4, 2/3, profile views accordingly and ground-truth. The blue bars in plots in the right column depicts the correlation distribution of correlation results.

Visual inspection showed that reconstruction results from the profile view were better than results from the 2/3 view, which was unexpected. This anomaly in results was also confirmed by the correlation values. There are several possible explanations to why this anomalous error appeared.

First, this anomaly in face reconstruction can be a result of weak face registration. The blurring in the frontal view of the surrogate multi-view image is evidence of poor registration of the different frames. From subchapter 3.3. we know that face registration is achieved by warping each remaining multi-view image from the chosen multi-view reference image. Hence, if a reference image will not be present in a neutral expression, with widely open eyes and the slightly open mouth, showing the teeth with a small black gap between them (so that the McGM can find a way to warp the reference image to reconstruct any dental or buccal features in the remaining images) (see chapter 3), then the face illustrations will not look nice and sharp. Thus, if face registration is weak and looks different after the warping process, the face alignment will also be affected, which may cause wrong reconstruction results.

Second, this anomaly could be caused by the poorer alignment of all views along the horizontal axis. The model developed in this thesis performs poorly if the dataset is misaligned. Even if data is perfectly aligned along the horizontal axis but is shifted up or down along the vertical axis, it will negatively affect mapping results.

Third, the presence of anomaly could be due to the face being in different shapes. A frontal view contains maximum information of the face, however, mapping a full face to profile did not have significant results, and reconstructed faces looked very poor. One way to explain it, that these differences may occur when both the driver and the target faces have different shapes, hence projecting the driver morph vector into the target PCA face space will create weird expressions on the target face. This mismatch may be improved or even avoided if we change the driver's morph vector components (e.g. warp and texture) before projecting it into the target face space. If the face features (eyes and philtrum) of a driver's face are not aligned sufficiently with features of the target face, an affine transformation should be applied. Next, they should be added to the target's morph mean vector because it will make a better alignment with the features of the target face.

Fourth, this anomaly could be more of a specific effect caused by something that occurred in the process for this particular set of data. It may be a generic process. Therefore, one may be required to go back to the beginning of the experiment and see if it is possible to improve the data set and then repeat the process with a corrected data set. Since this experiment was performed only once, it is difficult to know how the process would look like with a corrected or another dataset.

In the first part of this chapter's experiment, a unique PCA space was computed for each of the four asynchronous views of an individual’s face. The face was mapped from one view to another by taking components in the face-space of one perspective and mapping them into the face-space of another perspective. For this experiment, only four views of the face were used, e.g. frontal, 3/4, 2/3 and profile. In some cases, this strategy worked well. However, it appeared that better mapping between spaces could be obtained for views that were close to each other (e.g. frontal to 3/4 view, 18 degrees). When the views were too far apart, the process broke down to some degree (e.g. frontal to profile view, 54 degrees). It can be observed in the multi-view representation that was obtained from mapping using only frontal view face information. When the frontal face was mapped into 3/4 view, and the angle between perspectives was 18 degrees, the reconstructed face started to have small differences. The mouth area remained similar to the frontal face, while the eyes became slightly closed. A more prominent change in the reconstruction of the facial expressions was noticed after mapping to 2/3 or profile views. The eyes were nearly closed eyes, and the mouth became open. Thus, the face views that were placed quite close to each other brought better quality mapping results between PCA spaces than those who had wider angular differences. Similar results were obtained in Experiment 2 of chapter 4.

In the second part of the experiment, as it was mentioned earlier, a multi-view representation was chosen, obtained by mapping a series of frontal view representations into other views. It appeared to be the best multi-view representation among all the views. For this multi-view sequence, a PCA space was computed, and the object constancy technique for multi-views was applied (the technique is introduced in chapter 4).

6.3. Chapter 6 summary

The question addressed in this chapter is, for the most part, a cognitive science issue. Human observers do not have the opportunity of seeing multiple views at the same time. For that matter, we do not have the possibility of seeing a face in various illuminances or views at the same time. Therefore, the previous experiment didn’t appropriately reflect the way in which the human brain achieves object constancy for unfamiliar faces. Thus, multi-view representations were modified to explain how humans match unfamiliar faces. Here a multi-view representation was used, which was built from the experience of seeing asynchronous faces. The aim was to conduct a computational experiment to see if a multi-view vector can be created from asynchronous data and whether it could handle unfamiliar face reconstruction at vast angular differences between the viewpoints. The experiment strategy was to make separate PCA spaces for all four views and then map difference vectors from one perspective into these separate PCA spaces to find an image that would best represent the data in the different pose from the input face. Next, this strategy was evaluated by analysing the system (Experiment 3 of chapter 4). The results showed that when the face views were placed quite close to each other, good mapping quality could be achieved between PCA spaces (Experiment 2 of chapter 4). When the angle between the cameras increased to some degree, the mapping process started to break down. Hence, the mapping between PCA spaces is less productive when the angle between views is big.

7. Conclusions

In this thesis, we developed a set of new computational approaches for face reconstruction over multiple views, various illuminance conditions and asynchronous data sets. These approaches aimed to explain object constancy for faces and exploited a photo-realistic PCA-based facial model, which allowed to generate human facial actions from video footage of a moving face. This chapter will present a summary and discussion of the thesis and reveal the essential advantages and disadvantages of these new approaches. It will address the contribution this work made to the current state of the art and introduce future research directions for extending and improving current work.

7.1. Summary and discussion of the thesis

Object constancy for faces in the human visual system draws a great deal of interest, mainly because of its many applications in the various fields, including psychology, security, computer technology, medicine, computer graphics. Over the last decades, these fields have implemented multiple research studies and introduced various theories and approaches that tried to explain how to achieve object constancy for faces. Still, despite significant interest in this topic, they could not come to one successful and definite answer. For this reason, the current thesis aimed to develop a new PCA-based model approach of multiple appearances to explain object constancy for faces in human vision. This aim was achieved by creating three empirical pieces of research (chapter 4, 5, and 6), where a PCA-based multi-view mapping approach addressed the problem of reconstructing face under various pose, illuminance conditions and for unfamiliar faces. These pieces of research answered whether this approach is capable of replicating the built-in processes of the human visual system and can be used as a model of how humans achieve object constancy for faces.

First, to replicate integrated processes of the human brain, this thesis obtained its knowledge from the studies in psychology and cognitive science that focus on the human visual system. Chapter 1 provided a brief description of the face processing pathway and the vital part of the brain that is responsible for face perception, which is called a Fusiform Face Area (FFA). Also, it showed that the human visual system is viewpoint and lighting dependant, thus face constancy in the human brain is achieved in other ways. For instance, the systematic learning experience of changes in viewpoint directions or seeing faces in non-rigid facial motion can have a high impact on the process of face recognition and will help to achieve viewpoint invariance. To achieve lighting invariance, a face needs to be lit only by one light source positioned approximately over the head to recover simple shape from shading patterns. Familiarity with the 3-D depth structure of a face is equally important to obtain face constancy under various lighting conditions because face constancy is achieved by human vision processing faces holistically rather than based on the single features on the face. Most importantly, why this thesis started to favour and chosen a view-based approach to build a PCA-based model of multiple appearances is that the brain is storing face frames in the two-dimensional face space. Hence a 2D image-based approach was used to encode facial movements.

This thesis was setting up a computational model of a face to explain object constancy for faces, therefore, chapter 2 introduced the main strategies for encoding and animating faces in computer graphics together with the methods integrated into a new approach of this thesis and investigated how these methods relate to psychological theory on encoding the dynamic change in faces in the brain. In order to address the weaknesses of computer graphics algorithms, this study incorporated methods based on natural processes that are happening in the human brain. One of the methods is Principal Component Analysis (PCA). It is a well-known statistical technique that, while working on holistic information of faces, retains face information in 2D. This method was working as a content addressable memory used to recover information that is stored in the PCA space. This technique in more details was discussed in Chapter 3. Another method utilised in this thesis is Multi-channel Gradient Model (McGM), which is biologically motivated and introduced how motion perception is processed in human vision. It is an optic flow algorithm that was used to track relatively small facial motion in the video sequences.

Chapter 3 gave a detailed explanation of the hardware and software together with fundamental methods that were used to build a face system for this thesis.

Chapter 4 introduced empirical research to solve object consistency over multiple views, which was built on the existing PCA-based model introduced in chapter 3. The first experiment of this research showed that the developed technique not only succeeds in the performance-driven generation and animation of convincing photorealistic avatars, it also can distinguish between the rigid motion of the head and non-rigid movements in the face region over multiple-view sequence (Experiment 1, chapter 4). It showed that human vision could separate rigid motion of the head and non-rigid facial behaviour (expressions), which is very important for accurate facial recognition on the basis of dynamic image statistics. The second experiment showed that this technique is capable of reconstructing the face in one perspective from another perspective with high accuracy. This experiment allowed some insight into how the human visual system might deal with recognising faces in different independent perspectives, and the results suggest that people do not need to learn and store multiple face images from all the different perspectives in their head to accurately relate different views of the face (Experiment 2, chapter 4). Finally, the third experiment, the reproduction of missing viewpoints in a multi-view setup, produced an image of equivalent precision to the ground-truth (Experiment 3, chapter 4). In this experiment, a PCA space worked as a content addressable memory where all views were present. Having information only of one perspective, we reconstructed the missing viewpoints in a multi-view image. Initially, the reconstructed changes were muted. The quality of reconstruction has been improved by finding a scale factor, which defined the degree to which it needed to increase the weights of the reconstructed data to match the weights of the ground-truth data. After applying the scaling factor, the reconstructed data started to have a better match to rigid and non-rigid facial data. This method of recognising/reconstructing faces from different perspectives with all but one missing viewpoints makes use of a simple statistical model that could be used in the human visual system. These three experiments showed that the set of re-purposed existing techniques for generating and driving photo-realistic generated faces is a very powerful method for reconstructing and reproducing non-rigid facial information over multiple views.

In chapter 5, empirical research explored whether the method that was used to generate invariance over the pose in chapter 4 could work to explain lighting invariance. The experiment was performed over four different illuminance conditions, where existing techniques for generating and driving photo-realistic faces were re-purposed for reconstructing and reproducing non-rigid facial information over multiple lighting conditions. Every facial motion sequence was captured as a multi-illuminant image, consisting of three different illuminants. The multi-view reconstruction strategy has been applied to analyse multi-illuminant images. From a visual inspection, the reconstructed multi-illuminant images from the uniform lighting condition looked very similar to the original data representation before the scaling factor was applied. Results showed that the best match between the reconstructed and ground-truth data was obtained from data that was strongest illuminated in four different illuminance conditions. When the illuminant was weak, the reconstruction results were poorer. The initial results showed that rigid facial data provided a good match from all the illuminants, but non-rigid facial data only from the data with intense illuminant. Nevertheless, when the muted reconstruction of the facial behaviour has been amplified for different illuminants by scaling against ground-truth for all the reconstructions, overall illuminance conditions became successful. It has been identified that this general approach can relate faces captured under different illumination conditions. This method gave us an opportunity to answer how human vision is able to recognise/match/reconstruct faces over different lighting conditions.

Empirical research in chapter 6 created a computational experiment to show if a multi-view vector can be built from asynchronous data to explain face constancy for unfamiliar faces. A multi-view vector from asynchronous data was built in the laboratory settings. This experiment was divided into two computational parts, where in the first part, unique PCA spaces were computed for each of the four perspectives of an individual's face. These perspectives were unrelated to other different views. A face was mapped from one view to another by taking a difference vector (frame subtracted from the mean) in the face-space of one perspective and projecting them into the face-space of another perspective. Here, the computational method introduced in chapter 4 in experiment 2 has been replicated. In the second part of the experiment, the surrogate multi-view was composed of an image obtained from a frontal view from the first part of the computational experiment. For this multi-view sequence, a PCA space has been created, and the object constancy technique for multi-views been applied that as introduced in Experiment 3 of chapter 4. Here similar results were expected as in previous experiments using synchronous data. However, it appeared that even after finding and applying a scaling factor to match the original data, the best match was made from the frontal view representation. The images reconstructed from 3/4, 2/3 and profile views gave much poorer face reconstruction results. It would appear that the full face contains limited information about the profile view, hence face reconstruction in such views is low. The 3/4 view has information about the 2/3 view and frontal view, however, it did not give a good match to the ground-truth. Also, face reconstruction results obtained from the profile view were better than the 2/3 view. These results were unexpected. The possible reasons that could have caused this anomaly in results were also presented in chapter 6. Nevertheless, this work offered a proof of concept that the construction of a multi-view vector from asynchronous data is feasible. Therefore, a proposed approach may stand as a putative and alternative model of view invariance and lighting invariance in the human visual system.

7.2. Discussion of contributions

There have been many computational methods developed that tried to solve object/face reconstruction in various poses and lighting using different computer graphics strategies. For instance, the polygons approach. This approach could realistically replicate the anatomical structure of a face and facial dynamics, which made it a primary method for face reconstruction in the movie industry. Or view-based approaches that used only two-dimensional image information as a keyframe to construct a three dimensional morphable model of a face. But all of them had some disadvantages and could not address face reconstruction fully and invariant to wide poses and lighting changes. The approach developed in this thesis is capable to replace these complex face models. The model we developed does not use large databases or any landmark annotations to realistically reconstruct even small details on the face in the vast-angle between the viewpoints, while it replicates all the mechanisms that are built-in in the human visual system. Also, this model processes faces holistically and includes all the shading information of the face to achieve face constancy under various lighting conditions, the same as the human vision system does.

The evaluation of the performance of the PCA methods has been accomplished by comparing it with another image-based method, called morphing. Results showed that both techniques had proven to bring good results when the angle between views was small (around 5 degrees). However, the results of the morphing method failed badly when the angle between views increased up to 25 degrees, and only the PCA mapping approach with minor defects were able to obtain high reconstruction results. A multi-view method introduced in 4.1.5, on the other hand, can handle realistic faces reconstructions even when the angle between poses reaches 90-degrees. Comparison of the results showed that the morphing method would be working within a few degrees of rotation, but it will fail badly with a vast difference in face pose when the PCA multi-view approach is able to bring high face reconstruction despite a big change in perspective. Overall, the combination of performance-driven PCA-based mimicry and multi-view approach produced good results in the retrieval of missing viewpoints, illuminated faces and asynchronous faces from a multi-view facial representation.

7.3. Future directions

This thesis developed an image-based method for vectorising faces to explain object constancy for faces in three uncontrolled conditions. Experiments showed that this methodology was successful in face reconstruction over multiple poses and illuminations conditions, and to some degree, over asynchronous view-based data. Thus, future research on this topic should investigate what can be changed to improve the multi-view approach performance. One way to approach this would be to improve the computational aspect of this work. As mentioned in chapter 6, in the process of building a multi-view vector, there were some issues with particular views like a profile or 2/3 views. This process could be improved in several ways: by correcting alignment in the images, improving image registration, more data collection and testing on different datasets, adjusting pairs of views or perhaps a finer sampling of the views. For instance, a finer sampling of the profile view would make sense because if the change is made from the full-face to the face rotated by 30 degrees in the horizontal plane, it will not lead to a significant change in the face representation. However, if the degree of viewing angle will change from a 70-degree to a 90-degree angle, it will lead to a big difference in facial appearance.

Another aspect that should be explored in future research is empirical evidence, whether a better module of a face can influence an easier transfer from one view to another. As we know from experimental work in humans that it is quite hard to make an adjustment if the face is the same or not over different viewpoints. People do not recognise a face so easily from a different perspective. If they learn faces, they learn the face in one view and test in another. As we know from a study of Tamara Watson, view invariance is much better with dynamic information than with static information (Watson et al., 2005). Probably people can see the pattern of the continuous motion, and that carries over to profile better than just a shape of the face (an anomaly in results introduced in chapter 6).

When we are building these PCA spaces, it is basically an expression space, so we are using many videos of dynamic changes in the face. We are creating separate modules based on those dynamic changes and are looking at how the face changes for the viewpoints. Although people have shown (Hill & Johnston, 2001) that it is hard to recognise the face in different views, however maybe with a better module of the face, it would be easier to do? Finally, future research should investigate whether there are any more powerful and multipurpose vectorisation methods than those proposed in this thesis that could be used to optimise mathematical computations to speed up the whole face reconstruction process or transform this method into a fully automatic approach.

References

Adini, Y., Moses, Y., & Ullman, S. (1997). Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 721–732. https://doi.org/10.1109/34.598229

Allison, T., Puce, A., Spencer, D. D., & McCarthy, G. (1999). Electrophysiological studies of human face perception. I: Potential generated in occipitotemporal cortex by face and non-face stimuli. *Cerebral Cortex*, *9*, 415–430. https://doi.org/10.1093/cercor/9.5.415

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: Face-blind from birth. In *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2005.02.011

Beier, T., & Neely, S. (1992). Feature-based image metamorphosis. *ACM SIGGRAPH Computer Graphics*, *26*, 35–42. https://doi.org/10.1145/142920.134003

Berisha, F. (2009). *Photorealistic retrieval of occluded facial information using a performance-driven face model*. UCL (University College London).

Berisha, F., Johnston, A., & McOwan, P. W. (2010). Identifying regions that carry the best information about global facial configurations. *Journal of Vision*, *10*(11), 27. https://doi.org/10.1167/10.11.27

Beymer, D., & Poggio, T. (1995). Face recognition from one example view. *Proceedings of IEEE International Conference on Computer Vision*, 500–507. https://doi.org/10.1109/ICCV.1995.466898

Bilalic, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many Faces of Expertise: Fusiform Face Area in Chess Experts and Novices. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.5727-10.2011

Bilalic, Merim, Grottenthaler, T., Nagele, T., & Lindig, T. (2016). The Faces in Radiological Images: Fusiform Face Area Supports Radiological Expertise. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhu272

Blanz, V., Romdhani, S., & Vetter, T. (2002). Face identification across different poses and illuminations with a 3D morphable model. *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002*, 202–207. https://doi.org/10.1109/AFGR.2002.1004155

Blanz, Volker, Grother, P., Phillips, P. J., & Vetter, T. (2005). Face recognition based on frontal views generated from non-frontal images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2*, 454–461. https://doi.org/10.1109/CVPR.2005.150

Blanz, Volker, & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(9), 1063–1074. https://doi.org/10.1109/TPAMI.2003.1227983

Blanz, Volker, & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*, 187–194. https://doi.org/10.1145/311535.311556

Booth, M. C., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *8*(6), 510–523. https://doi.org/10.1093/cercor/8.6.510

Braje, W., Kersten, D., Tarr, M., & Troje, N. (1998). Illumination effects in face recognition. *Psychobiology*, *26*(4), 371–380.

Braje, W. L. (2003). Illumination encoding in face recognition: Effect of position shift. *Journal of Vision*. https://doi.org/10.1167/3.2.4

Brand, D. D. (n.d.). *Human Eye Frames Per Second: Just how many frames can we see per second?*

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology*, *46*(2), 369–384. https://doi.org/10.1152/jn.1981.46.2.369

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*, 105–116.

Bruce, Vicki, Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218. https://doi.org/10.1037//1076-898X.7.3.207

Bruce, Vicki, Valentine, T., & Baddeley, A. (1987). The basis of the 3/4 view advantage in face recognition. *Applied Cognitive Psychology*, *1*(2), 109–120. https://doi.org/10.1002/acp.2350010204

Bruce, Vicki, & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruce, Vicki, & Young, A. (1998). *In the eye of the beholder: The science of face perception.* Oxford University Press.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(1), 60–64. https://doi.org/10.1073/pnas.89.1.60

Burriss, L., Powell, D. A., & White, J. (2007). Psychophysiological and subjective indices of emotion as a function of age and gender. *Cognition and Emotion*. https://doi.org/10.1080/02699930600562235

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*. https://doi.org/10.1111/j.2044-8295.2011.02039.x

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from Security Surveillance. *Psychological Science*, *10*(3), 243–248. https://doi.org/10.1111/1467-9280.00144

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208. https://doi.org/10.1016/S0042-6989(01)00002-5

Cameron, J. (2009). *Avatar Exclusive -Behind The Scenes (The Art of Performance Capture)*. Media Magik Entertainment.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-8*(6), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Carey, S., Schonen, S. De, & Ellis, H. D. (1992). Becoming a Face Expert [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. https://doi.org/10.1098/rstb.1992.0012

Cowe, G. (2003). *Example-based Generated Facial Mimicry by*. *January*.

D, D., D, M., & M, S. (1998). An anthropometric face model using variational techniques. *In SIGGRAPH*, 67–74. https://doi.org/10.1145/280814.280823

Dalrymple, K. A., Corrow, S., Yonas, A., & Duchaine, B. (2012). Developmental prosopagnosia in childhood. *Cognitive Neuropsychology*, *29*(5–6), 393–418. https://doi.org/10.1080/02643294.2012.722547

Dalton, P. (1993). The role of stimulus familiarity in context-dependent recognition. *Memory & Cognition*, *21*(2), 223–234. https://doi.org/10.3758/BF03202735

Davies, G., & Milne, A. (1982). Recognizing faces in and out of context. *Current Psychological Research*, *2*(1–3), 235–246. https://doi.org/10.1007/BF03186766

DeCarlo, D., & Metaxas, D. (1996). Integration of optical flow and deformable models with applications to human face shape and motion estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.1996.517079

DeCarlo, D., Metaxas, D., & Stone, M. (1998). An anthropometric face model using variational techniques. *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, 67–74. https://doi.org/10.1145/280814.280823

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*. https://doi.org/10.1523/jneurosci.04-08-02051.1984

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Dimberg, U., Thunberg, M., & Grunedal, S. (2002). Facial reactions to emotional stimuli: Automatically controlled emotional responses. *Cognition and Emotion*. https://doi.org/10.1080/02699930143000356

Ding, C., Choi, J., Tao, D., & Davis, L. S. (2016). Multi-Directional Multi-Level Dual-Cross Patterns for Robust Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(3), 518–531. https://doi.org/10.1109/TPAMI.2015.2462338

Eger, E., Schweinberger, S. R., Dolan, R. J., & Henson, R. N. (2005). Familiarity enhances invariance of face representations in human ventral visual cortex: FMRI evidence. *NeuroImage*, *26*(4), 1128–1139. https://doi.org/10.1016/j.neuroimage.2005.03.010

Ekman, P, & Friesen, W. V. (1976). *Pictures of Facial Affect*. Consulting psychologists Press.

Ekman, Paul, & Friesen, W. V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. In *Consulting Psychologists Press*.

Essa, I. A., & Pentland, A. P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/34.598232

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is “special” about face perception? *Psychological Review*, *105*(3), 482–498. https://doi.org/10.1037/0033-295x.105.3.482

Fischer, M., Ekenel, H. K., & Stiefelhagen, R. (2012). Analysis of partial least squares for pose-invariant face recognition. *2012 IEEE 5th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012*. https://doi.org/10.1109/BTAS.2012.6374597

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395. https://doi.org/10.1145/358669.358692

Gauthier, I., & Logothetis, N. K. (2000). Is Face Recognition Not So Unique After All? *Cognitive Neuropsychology*, *17*(1–3), 125–142. https://doi.org/10.1080/026432900380535

Giese, M. A., & Leopold, D. A. (2005). Physiologically inspired neural model for the encoding of face spaces. *Neurocomputing*, *65–66*, 93–101. https://doi.org/10.1016/j.neucom.2004.10.060

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience*, *7*(5), 555–562. https://doi.org/10.1038/nn1224

Guo, K. (2013). Size-invariant facial expression categorization and associated gaze allocation within social interaction space. *Perception*, *42*(10), 1027–1042. https://doi.org/10.1068/p7552

Hancock, P. J. B., Bruce, V., & Mike Burton, A. (2000). Recognition of unfamiliar faces. In *Trends in Cognitive Sciences*. https://doi.org/10.1016/S1364-6613(00)01519-9

Hartley, R., & Zisserman, A. (2003). In computervision Multiple View Geometry in Computer Vision. *Computer-Aided Design*, *16*(2), 672.

Hasselmo, M. E., Rolls, E. T., Baylis, G. C., & Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, *75*(2), 417–429. https://doi.org/10.1007/BF00247948

Hill, H., & Bruce, V. (1993). Independent effects of lighting, orientation, and stereopsis on the hollow-face illusion. *Perception*, *22*(8), 887–897. https://doi.org/10.1068/p220887

Hill, H., & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology*, *11*(11), 880–885. https://doi.org/10.1016/S0960-9822(01)00243-3

Hill, H, Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition*, *62*(2), 201–222. https://doi.org/10.1016/S0010-0277(96)00785-8

Hill, Harold, & Bruce, V. (1996). The effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 986–1004. https://doi.org/10.1037//0096-1523.22.4.986

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106-154.2.

Jackson, A. S., Bulat, A., Argyriou, V., & Tzimiropoulos, G. (2017). Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 1031–1039. https://doi.org/10.1109/ICCV.2017.117

Johnston, A., Hill, H., & Carman, N. (1992). Recognising faces: Effects of lighting direction, inversion, and brightness reversal. *Perception*. https://doi.org/10.1068/p210365

Johnston, Alan, McOwan, P. W., & Benton, C. P. (1999). Robust velocity computation from a biologically motivated model of motion perception. *Proceedings of the Royal Society B: Biological Sciences*, *266*(1418), 509–518. https://doi.org/10.1098/rspb.1999.0666

Kanade, T., & Yamada, A. (2003). Multi-subregion based probabilistic approach toward pose-invariant face recognition. *Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation, CIRA*, *2*, 954–959. https://doi.org/10.1109/CIRA.2003.1222308

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1098/Rstb.2006.1934

Kaufmann, J. M., & Schweinberger, S. R. (2004). Expression influences the recognition of familiar faces. *Perception*, *33*(4), 399–408. https://doi.org/10.1068/p5083

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J Neurophysiol.* https://doi.org/10.1152/jn.1994.71.3.856

Kobatake, E., Wang, G., & Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *Journal of Neurophysiology*, *80*(1), 324–330. https://doi.org/10.1152/jn.1998.80.1.324

Kohonen, T. (1986). Self-organization, memorization, and associative recall of sensory information by brain-like adaptive networks. *International Journal of Quantum Chemistry*, *30*(13 S), 209–221. https://doi.org/10.1002/qua.560300821

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition*, *27*(6), 974–985. https://doi.org/10.3758/BF03201228

Lee, Y., Matsumiya, K., & Wilson, H. R. (2006). Size-invariant but viewpoint-dependent representation of faces. *Vision Research*, *46*(12), 1901–1910. https://doi.org/10.1016/j.visres.2005.12.008

Lewis, M. B. (2004). Face-space-R: Towards a unified account of face recognition. In *Visual Cognition*. https://doi.org/10.1080/13506280344000194

Logothetis, N K, & Sheinberg, D. L. (1996). Visual Object Recognition. *Annual Review of Neuroscience*, *19*(1), 577–621. https://doi.org/10.1146/annurev.ne.19.030196.003045

Logothetis, Nikos K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552–563. https://doi.org/10.1016/S0960-9822(95)00108-4

Lowe, David G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Lowe, D.G. (1999). Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1150–1157 vol.2. https://doi.org/10.1109/ICCV.1999.790410

Lucey, S., & Chen, T. (2008). A viewpoint invariant, sparsely registered, patch based, face verifier. *International Journal of Computer Vision*, *80*(1), 58–71. https://doi.org/10.1007/s11263-007-0119-z

Marr, D., & Nishihara, H. K. (1978). Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. *Proceedings of the Royal Society B: Biological Sciences*, *200*(1140), 269–294. https://doi.org/10.1098/rspb.1978.0020

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences*, *109*(42), 17063–17068. https://doi.org/10.1073/pnas.1116333109

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876. https://doi.org/10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2008). Matching Faces to Photographs: Poor Performance in Eyewitness Memory (Without the Memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364–372. https://doi.org/10.1037/a0013464

Meissirel, C., Wikler, K. C., Chalupa, L. M., & Rakic, P. (1997). Early divergence of magnocellular and parvocellular functional subsystems in the embryonic primate visual system. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.94.11.5900

Memon, A., & Bruce, V. (1985). Context effects in episodic studies of verbal and facial memory: A review. In *Current Psychology* (Vol. 4, Issue 4, pp. 349–369). https://doi.org/10.1007/BF02686589

Miyashita, Y. (1993). Inferior temporal cortex: Where Visual Perception Meets Memory. *Annual Review of Neuroscience*, *16*(1), 245–263. https://doi.org/10.1146/annurev.ne.16.030193.001333

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology*, *11*(4), 219–226. https://doi.org/10.1037/h0083717

Muja, M., & Lowe, D. G. (2009). Fast Approximate Nearest Neighbors With Automatic Algorithm Configuration. *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, 331–340. https://doi.org/10.5220/0001787803310340

Natu, V., & O'Toole, A. J. (2011). The neural processing of familiar and unfamiliar faces: A review and synopsis. In *British Journal of Psychology* (Vol. 102, Issue 4, pp. 726–747). https://doi.org/10.1111/j.2044-8295.2011.02053.x

Niyogi, P., Girosi, F., & Poggio, T. (1998). Incorporating Prior Information in Machine Learning by Creating Virtual Examples. *Proceedings of the IEEE.*, *86*(11), 2196–2209. https://doi.org/10.1109/5.726787

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A*. https://doi.org/10.1364/josaa.10.000405

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, *38*(15–16), 2351–2363. https://doi.org/10.1016/S0042-6989(98)00042-X

O'Toole, Alice J. (2005). Psychological and Neural Perspectives on Human Face Recognition. In *Handbook of Face Recognition*. https://doi.org/10.1007/0-387-27257-7_16

Parke, F. I. (1972). Computer generated animation of faces. *Proceedings of the ACM Annual Conference - Volume 1*, 451–457. http://doi.acm.org/10.1145/800193.569955

Parvizi, J., Jacques, C., Foster, B. L., Withoft, N., Rangarajan, V., Weiner, K. S., & Grill-Spector, K. (2012). Electrical Stimulation of Human Fusiform Face-Selective Regions Distorts Face Perception. *Journal of Neuroscience*, *32*(43), 14915–14920. https://doi.org/10.1523/JNEUROSCI.2609-12.2012

Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space*. *Philosophical Magazine Series 6*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. https://doi.org/10.1098/rstb.1992.0003

Perrett, D. I., & Oram, M. W. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, *11*(6), 317–333. https://doi.org/10.1016/0262-8856(93)90011-5

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K., Benson, P. J., & Thomas, S. (1991). Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Experimental Brain Research*, *86*(1), 159–173. https://doi.org/10.1007/BF00231050

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., & Salesin, D. (1998). Synthesizing realistic facial expressions from photographs. *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques SIGGRAPH 98*, *2*(3), 75–84. https://doi.org/10.1145/280814.280825

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, *4*(1), 0151–0156. https://doi.org/10.1371/journal.pcbi.0040027

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263–266. https://doi.org/10.1038/343263a0

Poggio, Tomaso, & Ullman, S. (2013). Vision: Are models of object recognition catching up with the brain? *Annals of the New York Academy of Sciences*, *1305*(1), 72–82. https://doi.org/10.1111/nyas.12148

Polat, U., Mizobe, K., Pettet, M. W., Kasamatsu, T., & Norcia, A. M. (1998). Collinear stimuli regulate visual responses depending on cell’s contrast threshold. *Nature*, *391*(6667), 580–584. https://doi.org/10.1038/35372

Pourtois, G., Schwartz, S., Seghier, M. L., Lazeyras, F., & Vuilleumier, P. (2005). View-independent coding of face identity in frontal and temporal cortices is modulated by familiarity: An event-related fMRI study. *NeuroImage*, *24*(4), 1214–1224. https://doi.org/10.1016/j.neuroimage.2004.10.038

Ramachandran, V. S. (1988). Perceiving shape from shading. *Scientific American*, *259*(2), 76–83. https://doi.org/10.3758/BF03206757

Rangarajan, V., Hermes, D., Foster, B. L., Weiner, K. S., Jacques, C., Grill-Spector, K., & Parvizi, J. (2014). Electrical stimulation of the left and right human fusiform gyrus causes different effects in conscious face perception. *The Journal of Neuroscience : The*

Official Journal of the Society for Neuroscience, *34*(38), 12828–12836. https://doi.org/10.1523/JNEUROSCI.0527-14.2014

Rezlescu, C., Pitcher, D., & Duchaine, B. (2012). Acquired prosopagnosia with spared within-class object recognition but impaired recognition of degraded basic-level objects. *Cognitive Neuropsychology*, *29*(4), 325–347. https://doi.org/10.1080/02643294.2012.749223

Rotshtein, P., Henson, R. N., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat Neurosci*, *8*(1), 107–113. https://doi.org/10.1038/nn1370

Russo, R., Ward, G., Geurts, H., & Scheres, A. (1999). When unfamiliarity matters: Changing environmental context between study and test affects recognition memory for unfamiliar stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*(2), 488–499. https://doi.org/10.1037/0278-7393.25.2.488

Rydfalk, M. (1987). *CANDIDE, a parameterized face, Report No. LiTH-ISY-I-866*.

Ryu, J. J., & Chaudhuri, A. (2006). Representations of familiar and unfamiliar faces as revealed by viewpoint-aftereffects. *Vision Research*, *46*(23), 4059–4063. https://doi.org/10.1016/j.visres.2006.07.018

Sergent, J., Ohta, S., & Macdonald, B. (1992). Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*. https://doi.org/10.1093/brain/115.1.15

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, *4*(3), 519. https://doi.org/10.1364/JOSAA.4.000519

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139. https://doi.org/10.1146/annurev.ne.19.030196.000545

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, *2*(1), 55–82. https://doi.org/10.3758/BF03214412

Terzopoulos, D., & Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(6), 569–579. https://doi.org/10.1109/34.216726

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, *9*(4), 483–484. https://doi.org/10.1068/p090483

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response Properties of the Human Fusiform Face Area. *Cognitive Neuropsychology*, *17*(1–3), 257–280. https://doi.org/10.1080/026432900380607

Tsai, R. Y. (1987). A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. *IEEE Journal on Robotics and Automation*, *3*(4), 323–344. https://doi.org/10.1109/JRA.1987.1087109

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience*, *31*, 411–437. https://doi.org/10.1146/annurev.neuro.30.051606.094238

Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

Ungerleider, L. G., & Haxby, J. V. (1994). “What” and “where” in the human brain. *Current Opinion in Neurobiology*. https://doi.org/10.1016/0959-4388(94)90066-3

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In *Analysis of Visual Behavior* (pp. 549–586). https://doi.org/10.2139/ssrn.1353746

Valentin, D., Abdi, H., O'Toole, A. J., & Cottrell, G. W. (1994). Connectionist models of face processing: A survey. *Pattern Recognition*. https://doi.org/10.1016/0031-3203(94)90006-X

Valentine, T. (1991). A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A*. https://doi.org/10.1080/14640749108400966

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology (2006)*, *69*(10), 1996–2019. https://doi.org/10.1080/17470218.2014.990392

Van der Linde, I., & Watson, T. (2010). A combinatorial study of pose effects in unfamiliar face recognition. *Vision Research*, *50*(5), 522–533. https://doi.org/10.1016/j.visres.2009.12.012

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, I-511-I–518. https://doi.org/10.1109/CVPR.2001.990517

Waters, K. (1987). A Muscle Model for Animating Three-Dimensional Facial Expression. *SIGGRAPH '87 Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, *21*(4), 17–24. https://doi.org/10.1145/37401.37405

Watson, T. L., Johnston, A., Hill, H. C. H., & Troje, N. F. (2005). Motion as a cue for viewpoint invariance. *Visual Cognition*, *12*(7), 1291–1308. https://doi.org/10.1080/13506280444000526

Weiner, K. S., & Grill-Spector, K. (2010). Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *NeuroImage*, *52*(4), 1559–1573. https://doi.org/10.1016/j.neuroimage.2010.04.262

Wikipedia contributors. (2019). *Avatar (2009 film)—{Wikipedia}{,} The Free Encyclopedia*.

Xu, Y. (2005). Revisiting the role of the fusiform face area in visual expertise. *Cerebral Cortex (New York, N.Y.: 1991)*, *15*(8), 1234–1242. https://doi.org/10.1093/cercor/bhi006

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. https://doi.org/10.1037/h0027474

Yin, R. K. (1970). Face recognition by brain-injured patients: A dissociable ability? *Neuropsychologia*, *8*(4), 395–402. https://doi.org/10.1016/0028-3932(70)90036-9

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330–1334. https://doi.org/10.1109/34.888718