A

Spott: on-the-spot e-commerce for television using deep learning based video analysis techniques

FLORIAN VANDECASTEELE, Ghent University - imec, ELIS - IDLab
KAREL VANDENBROUCKE, imec - MICT - Ghent University
DIMITRI SCHUURMAN, imec - MICT - Ghent University
STEVEN VERSTOCKT, Ghent University - imec, ELIS - IDLab

Spott is an innovative second screen mobile multimedia application which offers viewers relevant information on objects (e.g., clothing, furniture, food) they see and like on their television screens. The application enables interaction between TV audiences and brands, so producers and advertisers can offer potential consumers tailored promotions, e-shop items, and/or free samples. In line with the current views on innovation management, the technological excellence of the Spott application is coupled with iterative user involvement throughout the entire development process. This paper discusses both of these aspects and how they impact each other. First, we focus on the technological building blocks that facilitate the (semi-) automatic interactive tagging process of objects in the video streams. The majority of these building blocks extensively make use of novel and state-of-the-art deep learning concepts and methodologies. We show how these deep learning based video analysis techniques facilitate video summarization, semantic keyframe clustering and (similar) object retrieval. Secondly, we provide insights in user tests that have been performed to evaluate and optimize the application's user experience. The lessons learned from these open field tests have already been an essential input in the technology development and will further shape the future modifications to the Spott application.

Additional Key Words and Phrases: Interactive television; Video summarization; Deep learning; Object recognition; Metadata enrichment; Experience studies; User-validation

1. INTRODUCTION

With the explosion of mobile screens in the living room, TV as a medium has to cope with decreasing attention as viewers engage more and more in second screen activities [iMinds Digimeter 2014]. In fact, the use of second screen devices such as tablets and smart phones while watching television is increasing and due to this, the effective time of active watching is decreasing. On the other hand, there is a tendency of consumers who are inspired by the products they see on TV and subsequently want to buy these items. However, while mobile e-commerce is gaining more and more importance in the selling segment worldwide, TV viewers still struggle to find where to buy what their role model on TV is wearing or get to know the brand of furniture that is used in a scene. There are only a few tools available that link a limited amount of content on the screen with the e-commerce websites directly. Mostly, the user needs to search the desired object or product with a rough textual description or some brand information and in many cases, they fail to find the exact product.

The interactive second screen shopping platform Spott[1], which is discussed in this paper, tackles each of the issues mentioned above and allows users to instantaneously explore and buy what they see on television using their mobile second screen. The synchronization between both screens is performed using off-the-shelf audio recognition technology. This synchronization step tells us the program the user is watching and the exact video

[1]https://spott.it/en

This work is supported by the IWT/VLAIO O&O Spotshop project. The project's related e-commerce service - under the name 'Spott' - was launched on the market by project partner Appiness (20/04/16). Other project partners are Medialaan and BBDO. (More info: http://www.iminds.be/en/projects/spotshop).
Author's addresses: F. Vandecasteele and S. Verstockt, Ghent University - imec, ELIS - IDLab, Sint-Pietersnieuwstraat 41, B-9000 Ghent, florian.vandecasteele@ugent.be; K. Vandenbroucke and D. Schuurman, imec-MICT-Ghent University, Korte Meer 7-9-11, B-9000 Ghent.

shot that caused his action. The temporal metadata (i.e., product info, price, etc.) of the interactive objects in this video shot are generated with a novel (semi-)automatic video tagging solution, i.e., the deep learning based technological innovation this paper is mainly focusing on. The proposed video analysis techniques facilitate the tagging process and drastically reduce the manual labor of the annotation process by video summarization, object segmentation, class labeling, and semantically grouping similar objects and video frames.

Equally as important as the technological process behind the application is its user experience. For this reason, a lot of attention within the Spott project went to the real needs of users. Through our living lab approach, potential users are involved in each development stage of the application, i.e., the research is performed in a multi-disciplinary set-up where the end-user plays a central role. Results of the experience studies that were performed during development and launch of the application are included in this paper and show the potential of the application.

The remainder of this paper is organized as follows. Section 2 gives an overview of related platforms and applications. Subsequently, Section 3 presents the global workflow of the Spott application. Section 4 discusses the most important building blocks for video summarization and keyframe clustering and reduction. Both techniques decrease the computational costs and facilitate the tagging process. Next, Section 5 proposes our deep learning based solution for (similar) object retrieval in videos. Section 6 summarizes the experience studies and the adaption potential of the users which were placed central throughout this research. Finally, Section 7 lists the conclusions and points out directions for future work.

2. RELATED WORK

Over the last decade, we have witnessed the rise of a broad range of tools to find inspiration for clothing [Hadi Kiapour et al. 2015], e.g., Google Search, Pinterest (pinterest.com), Amazon flow[2]. Polyvore (polyvore.com), ASAP54 (asap54.com), Wheretoget (wheretoget.it), Pradux (pradux.com), WornOnTV (wornontv.net) and ShopYourTV (shopyourtv.com). Furthermore, several applications appeared on the market that allow television viewers to scan the content they are watching on TV and to detect the items they are searching for. Among these apps, the TheTake (thetake.com), SpyLight (spylight.com) and looklive (looklive.com) are the most popular ones. Although there seems to be some competition on the market, none of the above mentioned applications offers live interaction with television content, which is exactly the added value of Spott. Section 6 will show that this kind of interaction will change the experience of shopping and product placement. Contrarily to the existing apps, Spott also enables efficient detection of items in the broadcasted content and, most importantly, time efficient tagging of items in the background (with minimal manual intervention). The latter benefit is the biggest challenge for this kind of application since it takes a lot of time to manually tag products in every frame of a video. For this reason, our main technological focus was on transforming this process into a (semi-) automatic annotation chain, in which the human labeling task is strongly reduced.

3. FRAMEWORK

An overview of the Spott application, its technological innovations and user evaluation process is presented in Figure 1. In order to synchronize the TV screen with the mobile device the app starts recording and recognizing the sound of the active episode or soap on the television screen using an off-the-shelf audio fingerprinting technology. The idea behind

[2]https://www.a9.com/whatwedo/mobile-technology/flow-powered-by-amazon

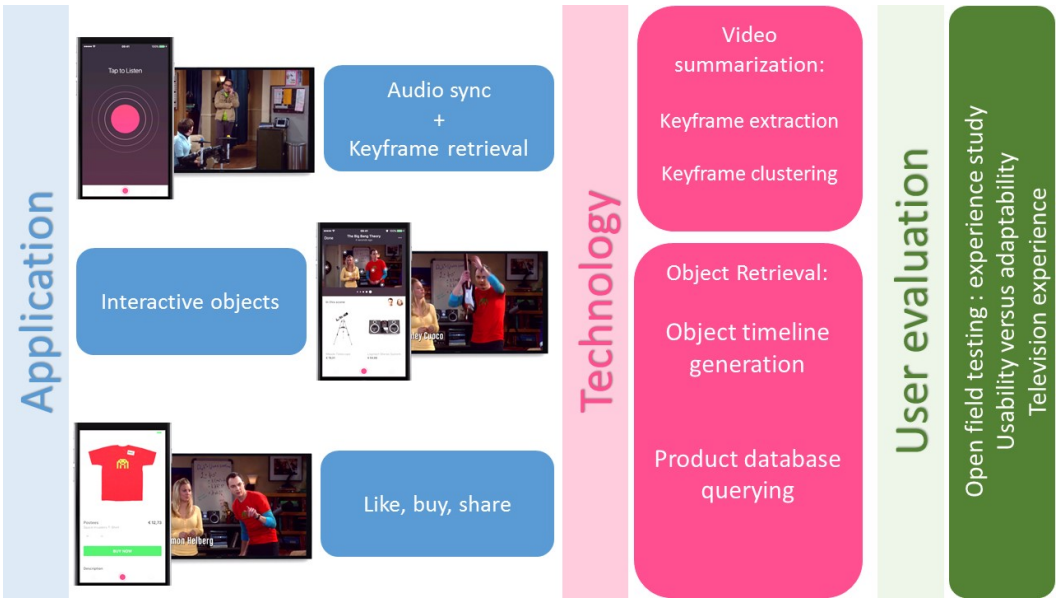

Fig. 1. Overview of the Spott application, its technological innovations and user evaluation process.

audio fingerprinting is the extraction of key properties such as tone, rhythm and tempo to create a unique fingerprint, feature vector for each video/ episode. This is done in the back-end for annotated video fragments and in the front-end the same extracting mechanism is used to create a fingerprint of the active episode. Subsequently, both fingerprints are matched and the temporal position of the video is retrieved. Several existing fingerprinting solutions have been compared and evaluated in varying set-ups with different levels of background noise to choose the most accurate solution within a maximum allowed time frame. Subsequently, based on the detected TV program and the location in the video stream, the corresponding keyframe is retrieved from our keyframe database and shown in the mobile app. The next section gives more details on the video summarization and keyframe generation process. In the next step, interactive objects are annotated in the frame as clickable dots.
For each of these dots/objects, a product image is given with some additional information. Finally, when a product is selected, the user can directly buy this item or like or share it with his friends. Within Figure 2 an overview is given of the back-end and the front-end pipeline. It is important to remark that some manual verification at the end of the back-end pipeline is still needed. A thorough description on how we transform the video keyframes into interactive objects is given in Section 5.1. The end user driven process in which all user actions have been designed and evaluated is discussed in Section 6.

4. VIDEO SUMMARIZATION

In order to optimize the Spott tagging process and decrease the computational cost of subsequent video processing tasks, it is important to reduce the amount of video data by filtering out redundant and unnecessary frames, while preserving only those frames, distinctive and essential to capture the entire video content. Furthermore, presenting the end-user with a limited list of representative keyframes improves his exploration and search process. Related to this, user evaluations have shown that keyframe based browsing of TV content has a positive impact on the interaction experience. The automated summarization of video content into representative keyframes, however, is a challenging problem due to

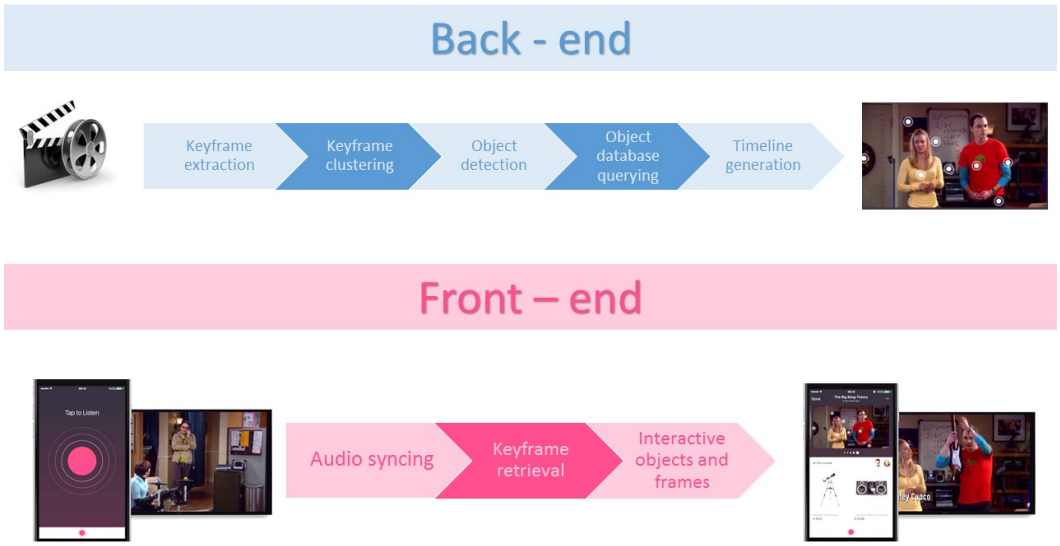

Fig. 2. Overview of the technical pipeline of the Spott application

the rapid change in lightning, viewpoint and scene during an episode. In order to cope with these issues we present a novel solution to extract, cluster and filter meaningful keyframes. This summarization process consists of three major steps.

(1) Keyframes are extracted from the video sequence with a histogram grid-based extraction mechanism.
(2) The number of keyframes is reduced by clustering visual similar frames, while preserving as much as possible of the entire content.
(3) A limited set of representative keyframes is given to the object retrieval algorithm (described in Section 5).

4.1. Histogram grid-based keyframe extraction

A lot of research has been done in the area of video summarization (i.e., keyframe retrieval). Ajmal et al. [Ajmal et al. 2012] give an overview of the different techniques and classification methods that are commonly discussed in literature, i.e., feature classification [Wang and Ngo 2012], clustering [dos Santos Belo et al. 2016], shot selection [Baraldi et al. 2015] and trajectory analysis [Qiu et al. 2008]. In this paper the grid-based histogram selection approach proposed by Vandecasteele et al. [Vandecasteele et al. 2016] is used, which has shown to outperform the former algorithms on a variety of television content. The technique performs a local histogram analysis on a 5-by-5 grid in combination with a keyframe quality analysis mechanism. This method ensures a good performance with fast camera movements, zoom gestures, gradual shot transitions and similar scene discrimination.

Subsequently, to cope with gradual shots like blends and fades an additional temporal analysis is performed. If the temporal distance between two shots is too small, these shots are considered as the same transition. It is important to remark that the temporal distance threshold is dependent on the type of content, i.e., action movies will have a short temporal distance, whereas romantic movies mostly have a larger temporal distance.

Finally, a weighted set of three no-reference quality metrics, that are suitable for real-time image quality analysis, are used to get the frame with the highest quality within the detected shot boundaries. The no-reference exposure metric, shown in Equation 1, is based on the average frame intensity and results in a value ranging from -1 to 1. A value of zero means that the image is correctly exposed. An exposure of -1 corresponds to underexposure and +1 represents an overexposed image.

$$exposure = \frac{\bar{x} - 127}{127}, \tag{1}$$

where $\bar{x}$ represents the average value of the histogram of the frame.

The contrast (Equation 2) on the other hand is based on the normalized spread of the histogram of the frame. This results in a value of 0 for images without contrast and 1 for keyframes with a maximal amount of contrast.

$$contrast = \frac{1}{128}\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2} \quad and \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{2}$$

where N represent the total amount of pixels in the frame and x_i represents the i-th pixel value of the frame.

Finally the sharpness (Equation 4.1) results in a value between -1 and 1, and is based on the fact that neutral lightning results in a histogram with all values centered around the middle of the keyframe histogram.

$$sharpness = \frac{1}{2 * 255^2 * Width * Height} * \tag{3}$$

$$\sum_{i=1}^{Width}\sum_{j=1}^{Height}[(\frac{\delta G(i,j)}{\delta i})^2 + (\frac{\delta G(i,j)}{\delta j})^2], \tag{4}$$

where $G(i, j)$ represent the pixel value of the frame.

The no-reference exposure, contrast and sharpness metrics are evaluated on a variety of television programs and are proven to be suitable for the intended purpose. Objective evaluation of our keyframe extraction algorithm on a manually generated ground truth dataset consisting of 400 video shots of different kinds of video content results in a recall of 99% and a precision of 89,9%. Different parts of the proposed summarization algorithm will be further investigated and optimized to decrease the number of false positive shot boundaries. The temporal distance threshold for shot boundaries, for example, could be automatically estimated based on the type of video content (and learned from end-user feedback). Finally, the amount of objects in the scene could also be used to select the best frame. Frames with more objects are more interesting for advertisers, but also for the users of the application, since they have more products to interact with.

4.2. Keyframe clustering

In general, a video scene consists of many shots that are visually similar. In a conversation, for example, each time the camera is focusing on another person this is seen as a new shot. Similarly, an entire video sequence can contain several temporally distant yet visually-similar frames that are taken in the same scene setting. On the one hand, removing these redundant and unnecessary frames improves the summarization process. On the other hand, clustering the similar keyframes will decrease the time for object tagging, i.e.,

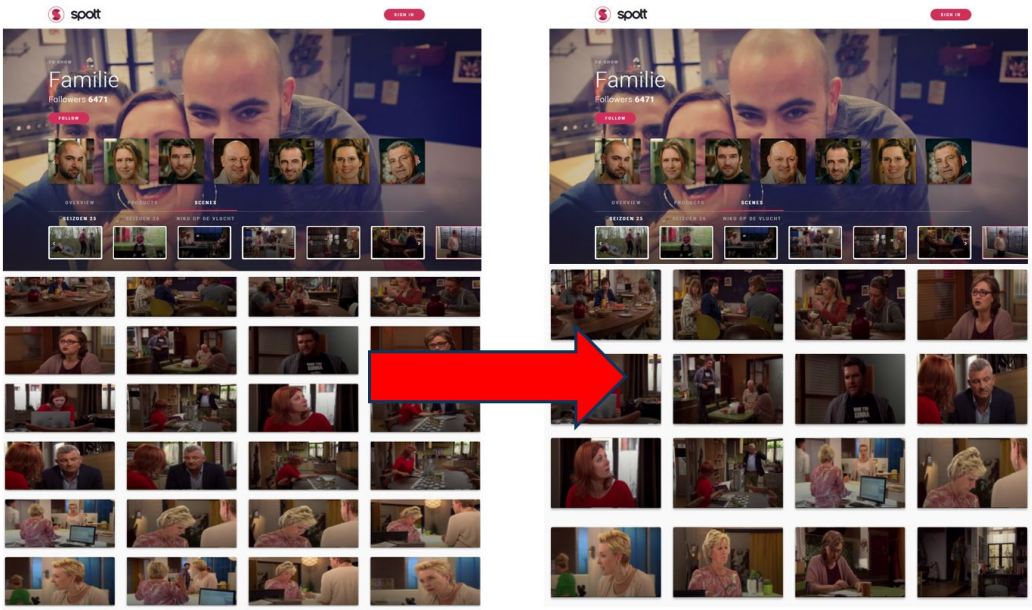

Fig. 3. Removing redundant keyframes to optimize the episode summarization and user interaction.

an automatic annotation can be performed on all object representations within the same cluster. Furthermore, it is interesting to summarize a complete video, episode to a limited set of images for representing the episode on a small screen or on a website (see Figure 3). User evaluations have pointed out that meaningful keyframe based browsing of TV content has a positive impact on the interaction experience.

The clustering process is done in three steps. First, we perform a global feature extraction with convolutional neural network (CNN) based learned features. Subsequently, we do a feature reduction by using the principal component analysis (PCA) technique. Finally, we use k-means clustering with the L2-distance of the reduced features of the keyframes.

The learning based features are generated by using a modified version of the Alexnet architecture [Krizhevsky et al. 2012] in which we removed the last 3 fully connected layers. The output after the last pooling layer is taken as the keyframe feature like proposed by Zagoruyko et al. [Zagoruyko and Komodakis 2015]. This ensures a higher level of image understanding and generalization (Zeiler et al. [Zeiler and Fergus 2014]) and avoids high visual similarity scores just based on global color or texture information. Furthermore, it is important to remark that if the amount of clusters is too high, redundant keyframes will appear. On the other hand, if the amount of clusters is too low, outliers (i.e., very unique keyframes or single scene shots) will not be shown. Figure 4 illustrates both of these remarks.

5. SIMILAR OBJECT RETRIEVAL

Once the representative keyframes of a video are detected, the goal is to link these with the shoppable content platforms.

Fig. 4. Keyframe clustering for a small and bigger number of clusters resulting in 5 and 40 keyframes respectively.

5.1. Video object timeline generation

To increase the object tagging and retrieval speed a fully automated mechanism is proposed to find similar objects/products in a set of keyframes. This makes it possible to spatio-temporally query different objects in a video sequence. The proposed object retrieval process consists of 3 major steps:

(1) We perform object detection and classification using the py-faster R-CNN [Ren et al. 2015] network. This network was retrained on the COCO dataset to improve the relevance of retrieved objects [Lin et al. 2014]. Furthermore, the non-relevant classes (i.e., classes of non-shoppable items: car, airplane, animals) are removed.

(2) For each detected object, a similarity score is calculated between the objects of the same class. This is done by evaluating the learning based features like proposed by Zagoruyko et al. [Zagoruyko and Komodakis 2015] and indicated in Section 4.2. Currently, the Euclidean distance between the two features is taken for the similarity score.

(3) Finally, the keyframes are spatially annotated with the object classes and their position to ensure a fast retrieval of similar objects. The major improvement of this process is that no manual input is required for drawing a bounding box around the object.

It is important to remark that the larger the set of objects in a particular class is, the slower the search retrieval process performs. However, detecting the objects can be performed in an offline process for non-live content. The major disadvantage of this approach is that those objects that are not part of the COCO dataset cannot be recognized and retrieved, consequently, for some keyframes the object detection framework fails to detect an appropriate number of objects. In [Tang et al. 2016], however, the authors propose a solution for this issue by means of transfer learning for weakly annotated or unknown classes. Figure 5 shows some examples of the proposed spatio-temporal object retrieval process, which consists of the following steps:

(1) The objects are detected, annotated and indexed in the current keyframe.
(2) An object is selected (book, bottle or wine glass) by clicking on the object and subsequently the most similar objects in the video sequence are searched.
(3) The location of the most similar objects, together with their keyframe timestamp, are shown in a similar object/keyframe list.

Currently, the evaluation in our application is done subjectively, but the data of the manual verification step will be used in the future to validate and improve our approach.

5.2. Clothing product database querying

The previous section described the linking of similar items throughout a video. However, there is no matching yet with the e-commerce websites or online shopping platforms. Especially for actors clothing, there were major interest to buy items from the screen. User-evaluations [Vandenbroucke and Schuurman 2016] pointed out that 71,8 percent wanted to buy casual clothing from Flemish soaps and 63,2 percent wanted to buy items while watching American sitcoms.

Although, the research community has actively been investigating the automated annotation and detection of clothes in digital content, there is currently no tool available that fully solves this problem. This is especially due to the large visual differences (see Figure 6) between the images retrieved on online webshops (with clean, white background, clear lighting conditions and a frontal view) and the clothing parts in a keyframe (i.e., cluttered background, unclear lighting conditions, side views, occluded body parts). To semi-automate the clothing linking we propose the spatio-temporal wardrobe generation architecture shown in Figure 7. The current framework consist of the following 4 building blocks:

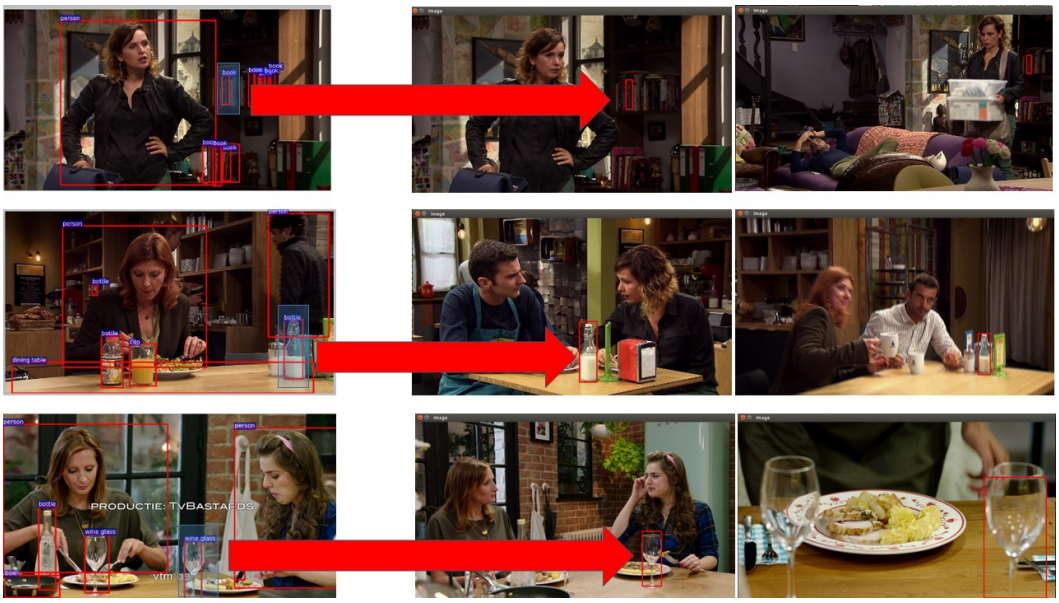

Fig. 5. Spatio-temporal similar object retrieval in keyframes.

Fig. 6. Visual differences between clothing images of an online webshop and clothing items in a keyframe.

(1) A face detection and recognition algorithm searches the faces in the image and annotates them with facial metadata (such as gender and age, which can be used as filters on the shopping platforms).
(2) Based on the face position, we perform a clothing object segmentation.
(3) With the selected clothing object we perform a clothing recognition.
(4) Finally, we perform coarse-to-fine clothing matching with an e-commerce annotated dataset.

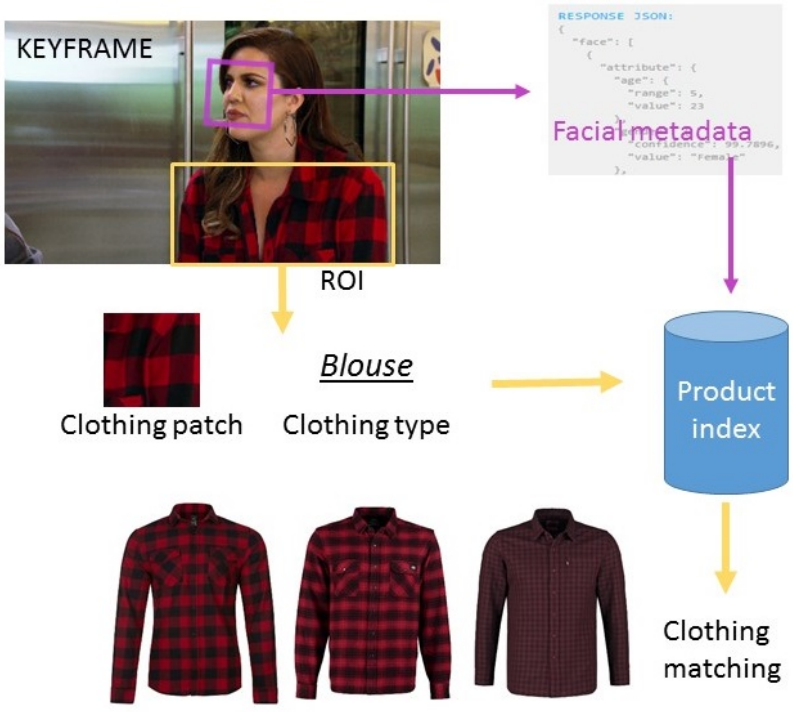

Fig. 7. Spatio-temporal wardrobe generation framework.

Our algorithm starts with face detection. A positive face detection ensures that there is a person in the keyframe and with the face pose estimation it is possible to localize the corpus. The facial features itself are used for generating the gender and age metadata filters. Important to remark is that keyframes without a face are removed, which drastically reduces the set of keyframes. Since the main focus of this paper (and the Spott project in general) is not on the improvement of the current available frameworks for face detection and recognition, a commercial off-the-shelf algorithm is used.

The second part of our algorithm is the clothing patch region-of-interest (ROI) selection. This step is based on the facial information (i.e., the position, the rotation and dimension) and the position of the lowest neck pixels (clothing boundary). The ROI patch is taken below the lowest neck pixel, taking into account the face orientation. From this patch, we use the color and pattern information to query the clothing database. It is important to remark, however, that this step could fail in case there are occluded body parts or different clothing parts on top of each other.

The third element in our framework is the clothing type recognition (i.e., shirt, blouse, T-shirt) which is done by fine-tuning the pre-trained VGG (Visual Geometry Group) network of Simonyan et al. on a dataset containing 7210 samples of over 18 different clothing element classes. Basic preprocessing steps such as subtracting the mean image values and generating randomly cropped, mirrored and shifted images are used to augment our dataset. Furthermore, the finetuning is done by using minibatch gradient descent in combination with momentum.
The evaluation of this technique is done on keyframes from different television shows and

the network achieves a top-1 accuracy of 35% and a top-3 accuracy of 47%. It is important to remark that the training data that we used is based on non-occluded clothing images, whereas the testing keyframes contain occluded clothing parts. As indicated in previous sections, training on the current manual verified data will help in increasing our final accuracy to be comparable with state-of-the-art approaches [Hadi Kiapour et al. 2015; Wang et al. 2015].

The final step of our framework is the coarse-to-fine linking. The gender and age information from the face and the clothing type are used as an initial filter on the semantic annotated clothing dataset. Subsequently, a query with the CEDD [Chatzichristofis and Boutalis 2008] and PHOG [Bosch et al. 2007] features from the patch are generated on the filtered clothing dataset. Furthermore, some additional filters could be added such as an actor-brand relationship. Finally, our lead users indicated subjectively that the top-3 matches are highly relevant.

Currently, the proposed methodology for object linking with e-commerce websites is only evaluated for clothing on an online retrieved dataset from Zalando, but the proposed coarse-to-fine strategy will also work with other types of objects. Our algorithm that uses color and texture information along with meta-data such as an object type or brand name is a generic methodology that only needs trained classifiers for a specific task. Furthermore, the same procedure could also been used to find relevant or related attributes given a specific object.

6. EXPERIENCE STUDIES

6.1. Living Lab research methodology

Technological excellence is not the only factor that drives market success, innovations should also tap into the end-users' unfulfilled needs. Since the late 1990s, more and more attention has gone out to measuring and understanding how users experience ICT products, applications and services. One of the drivers in this respect is that User Experience (UX) has been linked to market success or failure of new and existing products, applications and services [De Marez and De Moor 2007].

To investigate the end-user experiences with Spott, a Living Lab research project, which consists of a multi-method, qualitative and quantitative mix of end-user research methods, and based on customer-led ideation and co-creation, was set up [Schuurman 2015]. One of the distinct characteristics of Living Labs research is that it gathers insights from the potential customer in a real-life context. Subsequently, the captured feedback is more reliable.

The Spott Living Lab research project consisted of 6 research steps conducted between June 2015 and July 2016: desk-research, co-creation with potential end-users (N=9), usability testing of prototype (N=12), closed field test (N=254), adoption potential estimation survey (N=307) and an open field test with more than 10.000 users. A full overview of the research track is described by Vandenbroucke et al. [Vandenbroucke and Schuurman 2016]. In this section, the main outcomes of the last research step, i.e., the open field test, will be discussed.

6.2. Open field test with adoption potential measurement

After the application was developed and thoroughly tested in a closed field test, Spott was launched in both the iTunes Store and the Google Play Store. By means of a marketing campaign, consisting of press releases, mailing and paid advertisements, the app was introduced to the Flemish television audience in April 2016. In May 2016, a call to participate in the open field test was sent out by iMinds Living Labs and Medialaan. Concretely, over 60.000 people received an e-mail in which they were asked to download the application and provide their feedback in 3 online surveys that were distributed in June and July 2016. Completing all three surveys was not mandatory. People that did not download or use the application were also asked to provide their feedback on specific questions of the survey.

The focus of the first survey was on the usability of the app measured by means of the System Usability Scale (SUS) [Brooke et al. 1996] The second survey focused on TV experience and the last survey focused on the potential of the app for advertisers. In all 3 surveys, the potential for usage (i.e., adoption potential for current non-users) and potential for repeated usage (for current users) was measured by means of the Product Specific Adoption Potential (PSAP) methodology proposed by De Marez et al. [De Marez and Verleye 2004], together with use frequency of the application and socio-demographic information.

With the PSAP methodology, respondents are assigned to Rogers' innovation profiles (i.e., innovators, early adopters, early majority, late majority and laggards) [Rogers 2010] based on their answers to 3 statements (5-point Likert-scales;(1) definitely use (5) definitely not use). The first statement measures the degree to which the participant would use the application. The second, optimal statement measures the degree to which the participant would use the application with specific features. The third, sub-optimal statement measures the degree to which the participant would use the application without the specific features. The specific features in the first survey were the national and international content with which they would preferably use the application. In the second survey, these features are related to price comparison between web shops and buying similar products as the recognized products in the app. In the third survey, these features are related to buying the products in the application and getting information about where to buy the items in an offline store.

In total, 343, 460 and 204 respondents took part in respectively the first, second and third online survey. A cohort of 46 respondents completed all 3 surveys. The lower amount of respondents of the third survey is probably due to the summer holidays in Belgium and survey fatigue. The subjective survey results were linked with objective logged usage data of the application, which provides a detailed and better understanding of the applications actual usage and adoption potential. In the following subsections, insights will be provided about the apps usability, adoption potential and repeated usage.

6.3. Usability

In total, 160 participants completed the System Usability Scale, which resulted in a median score of 75/100 (minimum score: 32.5/100; maximum score: 100/100). When a median SUS score is higher than 68/100, usability is good. As can be seen in the figure below, 91.3 percent rated the application good, excellent or best imaginable. This is also noticeable by the fact that only 10.0 percent indicated that synchronizing with TV content didnt go fluent. When taking a look at the logdata, we saw that it took on average 2 seconds to

synchronize with TV-content, i.e., for the app to recognize and display the TV-content on the smart phone.

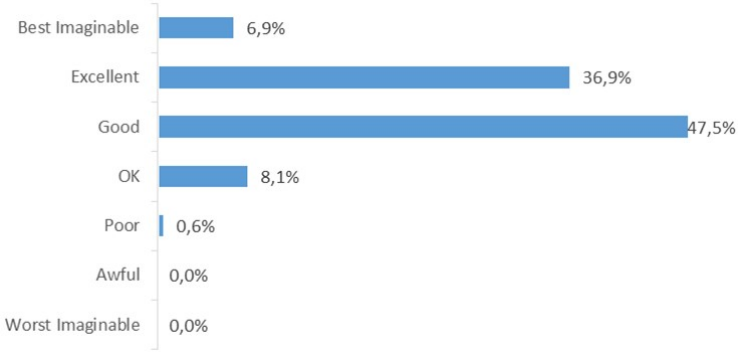

Fig. 8. Adjective UX rating scale (N=160).

6.4. Television experience

For a large part of the end-users, we measured a positive influence of the apps usage on their TV experience. In survey 2, 37.5 percent of 104 respondents reported that they feel more attentive and 41.3 percent reported that they feel more involved with TV content when using the application. 42.3 percent thinks that they have an lengthened TV experience and 69.2 percent agreed with the statement that Spott makes them more attentive with products they see on TV.

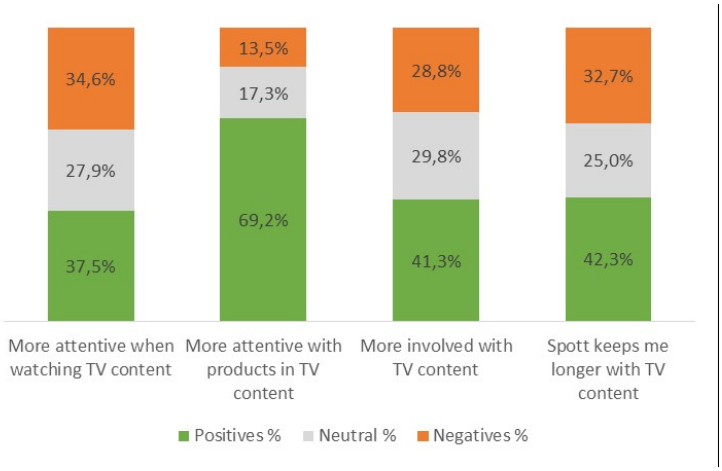

Fig. 9. Influence of Spott on TV experience (N=104).

6.5. Adoption potential measurement

As can be seen in Figure 10, we notice a high adoption and usage potential for the application. More than 70 percent of the Innovators and Early Adopters are female, with a mean age of 27.5 years (S.D. age= 13.6). The Laggard group is more masculine (41.2 percent females) and older (mean age = 41.5 years; S.D. age= 20.2). Over the course of the 3 surveys, there is a drop noticeable in adoption and usage potential. There are 4 explanations for this. A first reason for the high adoption potential measured in the first survey is that usage is mostly content driven. Participants indicated that they only want to use the application with the content they are watching, and not use the application with content they dont like. A second explanation is the fact that 65.6 percent agreed that the application is inspiring, while 47.5 percent agreed that the application is commercial (N=160). Thus, respondents were less prone to indicate they would use the app when they were confronted with the commercial features. A third explanation lies in the fact that the third survey was conducted during the summer holidays in Belgium, this is a period with less content on TV that was interactive. A fourth explanation can be found in the rather low repeated usage of the application, as was seen in the logdata, 30 days after the application was publicly launched in April, 61.3 percent of 9979 downloaders had used it once, 31.6 percent used it 2 or 3 times and 7.1 percent used it 4 times or more. This means that users are interested in testing the application, but they did not form the habit to use it on a regular basis yet. Nonetheless, it can be stated that, even in the third survey, we still measured a high adoption potential, i.e., there are more innovators and early adopters than in the theoretical Rogers curve.

	INNOVATORS	EARLY ADOPTERS	EARLY MAJORITY	LATE MAJORITY	LAGGARDS
ROGERS THEORY	2,5%	13,5%	34,0%	34,0%	16,0%
PSAP WAVE 1 (N = 343)	19,8%	46,1%	7,2%	15,9%	11,1%
PSAP WAVE 2 (N = 460)	13,0%	38,7%	20,4%	20,2%	7,6%
PSAP WAVE 3 (N = 198)	8,1%	26,3%	25,3%	27,8%	12,6%

Fig. 10. Application's usage potential compared to Rogers' Theory.

7. CONCLUSIONS AND FUTURE WORK

This paper presented the deep learning based architecture and experience study of an interactive second screen application which offers viewers relevant information on objects they see and like on their television screen. Different building blocks for the semi-automatic tagging process were described along with the user evaluation tests. The video summarization framework for keyframe extraction and clustering in combination with the object retrieval algorithm ensures a fast and semi-automated linking of shopping content. This allows for time efficient tagging of TV-content, which is currently a challenge for existing alternatives on the market.

From the user experience studies, it can be concluded that the application provides a good user and television experience for end-users. Moreover, we measured a high adoption potential, indicating that this application taps into an unfulfilled need which will change the current way of product placement. Therefore, this might be an interesting product for telco's and broadcasters to engage their audiences to spend more attention to the products used in their programs. Furthermore, our experience studies have shown that users also want to use the Spott application in an non-interacted modus (i.e., not synced with the TV) to find similar products or complementary items, which is also provided by our object retrieval algorithm. This non-interacted version of Spott will increase the involvement of the users with their favorite actors or episodes.

The proposed methodology drastically reduces the tagging time needed to annotate the most prominent objects in the video streams. However, some manual verification is still needed. Future work will further optimize the video content processing to ensure a fully automated object tagging pipeline. With a real-time tagging mechanism it should even be possible to annotate live streams in the near future. Further improvements using novel image analysis, big data and deep learning techniques will be implemented and fitted to our proposed methodology to enhance the current framework.

For the video summarization block, future work will consider to incorporate facial information to improve the selection of the best frame (in addition to the currently used no-reference metrics). Keyframes with frontal faces of the actors are more likely to be important compared to those with side views. This will also affect the clothing retrieval process which heavily relies on the meta-data retrieved from the face detector. Furthermore, we will evaluate several neural network architectures, such as VGG (Visual Geometry Group), for optimizing the automated keyframe feature extraction. In addition, the automated detection of representative clusters will be further investigated.

Within the video object timeline generation future research will focus on improved models and techniques for object detection and localization to increase the detection accuracy. Related to this, ensembles of different object detection models have already proven to increase the object localization and recognition task [Zagoruyko et al. 2016].

For the clothing product database querying, future research will be necessary to improve the clothing patch selection process. Conditional random fields as proposed by Serra et al. [Simo-Serra et al. 2014] or LSTM graphs [Liang et al. 2016] will be further evaluated in this context. Finally, for the clothing type recognition more evaluation and optimization will be done on recent datasets [Liu et al. 2016].

ACKNOWLEDGMENTS

The research activities as described in this paper were funded by Ghent University, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) and Appiness bvba. The authors would also like to thank the media partners Medialaan, BBDO for their willingness to share their content and finally also a special thank for the people that filled in the questionnaires. Without their help no decent user-evaluation would be possible.

REFERENCES

Muhammad Ajmal, Muhammad Husnain Ashraf, Muhammad Shakir, Yasir Abbas, and Faiz Ali Shah. 2012. Video summarization: techniques and classification. In *Computer Vision and Graphics*. Springer, 1–13.

Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2015. Shot and scene detection via hierarchical clustering for re-using broadcast video. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 801–811.

Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 401–408.

John Brooke and others. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

Savvas A Chatzichristofis and Yiannis S Boutalis. 2008. CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *International Conference on Computer Vision Systems*. Springer, 312–322.

Lieven De Marez and Katrien De Moor. 2007. The challenge of user-and QoE-centric research and product development in today's ICT-environment. *Observatorio (OBS*)* 1, 3 (2007).

Lieven De Marez and Gino Verleye. 2004. Innovation diffusion: The need for more accurate consumer insight. Illustration of the PSAP scale as a segmentation instrument. *Journal of Targeting, Measurement and Analysis for Marketing* 13, 1 (2004), 32–49.

Luciana dos Santos Belo, Carlos Antônio Caetano, Zenilton Kleber Gonçalves do Patrocínio, and Silvio Jamil Ferzoli Guimarães. 2016. Summarizing video sequence using a graph-based hierarchical approach. *Neurocomputing* (2016), 1001–1016.

M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE International Conference on Computer Vision*. 3343–3351.

iMinds Digimeter. 2014. Adoption and usage of media and ICT in Flanders. Research report, Ghent, iMinds. (2014). https://www.iminds.be/en/gain-insights/digimeter

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. 2016. Semantic Object Parsing with Graph LSTM. *arXiv preprint arXiv:1603.07063* (2016).

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

Kuan-Hsien Liu, Ting-Yen Chen, and Chu-Song Chen. 2016. MVC: A Dataset for View-Invariant Clothing Retrieval and Attribute Prediction. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 313–316.

Xuekan Qiu, Shuqiang Jiang, Huiying Liu, Qingming Huang, and Longbing Cao. 2008. Spatial-temporal attention analysis for home video. In *IEEE International Conference on Multimedia and Expo*. IEEE, 1517–1520.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

Everett M Rogers. 2010. *Diffusion of innovations*. Simon and Schuster.

Dimitri Schuurman. 2015. *Bridging the gap between Open and User Innovation?: exploring the value of Living Labs as a means to structure user contribution and manage distributed innovation*. Ph.D. Dissertation. Ghent University.

Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. 2014. A high performance CRF model for clothes parsing. In *Asian conference on computer vision*. Springer, 64–81.

Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandrea, Robert Gaizauskas, and Liming Chen. 2016. Large Scale Semi-Supervised Object Detection Using Visual and Semantic Knowledge Transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Florian Vandecasteele, Jeroen Vervaeke, Baptist Vandersmissen, Michel De Wachter, and Steven Verstockt. 2016. Spatio-Temporal Wardrobe Generation of Actors Clothing in Video Content. In *International Conference on Human-Computer Interaction*. Springer International Publishing, 448–459.

Karel Vandenbroucke and Dimitri Schuurman. 2016. APPTVATE: The mobile shopping technology that enriches your TV experienc. *iMinds-Appiness-Medialaan* (2016).

Feng Wang and Chong-Wah Ngo. 2012. Summarizing rushes videos by motion, object, and event understanding. *IEEE Transactions on Multimedia* (2012), 76–87.

Haoran Wang, Zhengzhong Zhou, Changcheng Xiao, and Liqing Zhang. 2015. Content based image search for clothing recommendations in e-commerce. In *Multimedia Data Mining and Analytics*. Springer, 253–267.

Sergey Zagoruyko and Nikos Komodakis. 2015. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4353–4361.

Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro O Pinheiro, Sam Gross, Soumith Chintala, and Piotr Dollár. 2016. A MultiPath Network for Object Detection. *arXiv preprint arXiv:1604.02135* (2016).

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

Received November 2016; revised March 2016; accepted June 2017