DiEvent: Towards an Automated Framework for Analyzing Dining Events

Mahmoud Qodseya, Mahdi Washha, Florence Sedes
IRIT - Paul Sabatier University
Toulouse, France
Mahmoud.Qodseya,Mahdi.washha,Florence.sedes@irit.fr

***Abstract*—The analysis of dining events is important and useful for a wide range of applications such as smart restaurants, and for different research areas like sociologists' studies and social interactions analysis. Particularly, the performance and reputation of restaurants is completely dependent on customers satisfaction as a possible metric. With the rapid growth of computer vision technologies, smart restaurants are going to leverage such technologies for indirectly measuring and quantifying customer satisfactions through analyzing videos and images without performing any direct questioner process. However, the large volume of recorded data by cameras makes the manual analysis process computationally expensive in terms of time to quantify customer satisfaction. Hence, in this paper, we introduce a design of a framework, so-called DiEvent, that integrates various components for automatically analyzing dining events. Our framework could be leveraged in different applications and researches, including cooking recipe evaluation in terms of customer satisfaction, performing sociology studies in dining events, and social interactions detection researches.**

I. Introduction

Dining activity is a common type of social events, in which people usually enjoy eating, drinking, and talking. Indeed, the analysis of dining activities has gained importance for many smart applications and research fields such as smart restaurants, gastronomy, social interactions detection and analysis [1], [2], [3], [4], and the evaluation of dining tools like Georgette[1]. Sociologists are also interested in the relation between emotion and eating, where the eating behavior of the human being is directly influenced by emotions and vice versa [5].

The main challenging problem in analyzing social events is the impracticality of doing that in a manual way because of the huge volume of data in terms of videos and images that need annotation. Therefore, such a serious problem raises the need for an automated tool/system that performs a deep analysis for recorded videos or captured images. Smart restaurants and sociologists are the main beneficiaries of having automated analysis dining events system. More precisely, smart restaurants can quantify their services quality throughout indirectly measuring customers satisfaction. For instance, cooking recipe evaluation can be indirectly measured by analysis customers' facial expression (e.g., happy, disgust). Also, such an automated system can facilitate the job of sociologist in many aspects: 1) providing deep details by analyzing multiple views of scenes; 2) enriching the scope of eating behavior with empirical datasets; 3) detecting and highlighting the most important scenes, shots, and events inside videos; 4) and reducing the time needed for analyzing a video by sociologists or locating the relevant scenes.

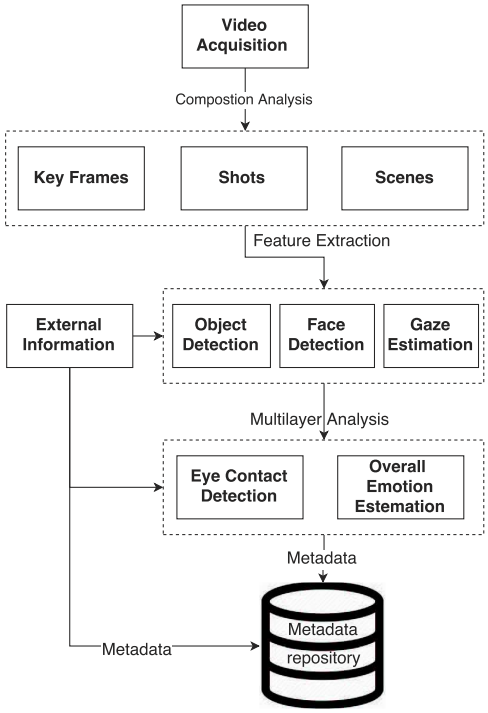


Fig. 1: DiEvent framework pipeline.

To build a framework that can automatically analyze dining social events, we need to integrate many individual components: 1) objects detection; 2) human face detection; 3) human face tracking; 4) facial expression detection; 5) human gaze estimation; 6) and video summarization. Human face and facial expression recognition are essential in many applications such as video surveillance [6], privacy preserving [7], and

[1]The 'Georgette', a spoon which is both a fork and a knife and designed by Jean-Louis Orengo. http://www.downtownmagazinenyc.com/the-georgette-the-evolution-of-the-spork/

DOI 10.1109/ICDEW.2018.00034

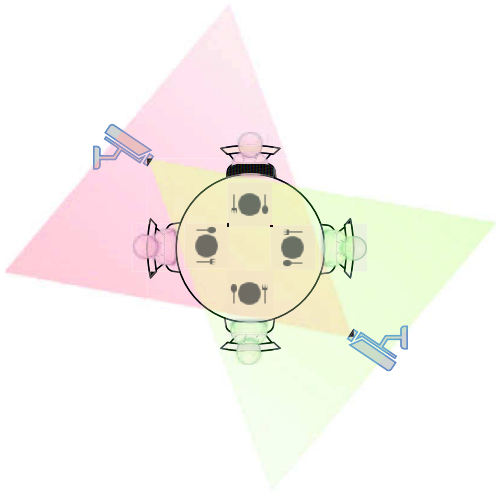


Fig. 2: Video acquisition schema. It includes two surveillance cameras fixed in front of each other at hight of 2.5 meters with $-15°$ pitch angle.

object detection and tracking [8]. Furthermore, the facial expressions represent a communication system among actors and observers, in which they exchange signals that can affect their social interactions [9]. However, there are many challenges in face detection such as $pose$ (position and orientation), facial expression, face occlusion, and imaging conditions [10].

Gaze direction has a crucial role in representing human visual selection and attention [11]. People look each other during the social interactions. Also, the eye-contact provides multiple functions in the two-person contacts [12] such as information seeking, establishment and recognition of social relationships, and signaling that the "channel is open".

Karpathy and Fei−Fei [13] present a model that matches language and visual data. This model is able to generate natural language descriptions of images and their regions. Similarly, Otani *et al.* [14] present a video summarization technique to provide an overview of the video contact. They use a deep neural network to map videos and descriptions to a common semantic space.

Codreanu *et al.* [15] propose a generic metadata model that gathers various features extracted from a video, detecting people, objects, events, *etc*. Gao *et al.* [16] introduce an algorithm to analyze dining activity in a nursing home using a Hidden Markov Model to determine the relevant dining event.

In this paper, we propose a framework $(DiEvent)$ for analyzing dining events. As shown in Figure 1, $DiEvent$ consists of five sequenced steps: 1) video acquisition platform; 2) video structure analysis to decompose the video into key frames, shots, and scenes; 3) feature extraction in which we use the OpenFace[2] toolkit [17] for face detection and gaze estimation, OpenFace[3] library for face recognition [18], and a trained model for emotion recognition; 4) extendable multilayer analysis to detect the eye contact and estimate the overall emotion; 5) and metadata repository to store both extracted and collected metadata for querying scenes $w.r.t.$ a particular context.

[2]OpenFace: an open source facial behavior analysis toolkit.

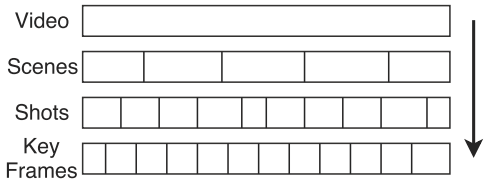


Fig. 3: Video parsing hierarchy [19].

The key points of using the acquisition platform are to: 1) build a dataset to analyze the social events; 2) collect more external information such as location, number of participants, temperature, social relationships, *etc*; 3) and have a wide view using multiple cameras.

II. DiEvent Framework Design

In this section, we introduce the methods that are necessary to analyze the dining social events. The DiEvent framework takes care of data acquisition, video composition analysis, feature extraction, multilayer analysis, and storing both collected and extracted metadata in a repository.

A. Video Acquisition Platform

Figure 2 shows the acquisition platform consisting of two surveillance cameras. The cameras are placed in front of each other at hight of 2.5 meters with $-15°$ pitch angle, to capture the corresponding parts of the scene. The cameras acquire images at a frame rate of 25 fps with a resolution of 640×480 pixels.

B. Video Composition Analysis

We apply video parsing technique to segment a video into structural parts as described in Figure 3. Video parsing process includes: 1) shot boundary detection; 2) key frame extraction; 3) and scene segmentation (see [19] for a review of video composition analysis).

C. Feature Extraction

Feature extraction represents a crucial preprocessing step since it transforms the high-dimensional space data to lower-dimensional space. For this purpose, we use OpenFace toolkit, trained model for emotion recognition, and OpenFace library for face recognition.

OpenFace toolkit is an open source framework that implements state-of-the-art facial behavior analysis algorithms, including facial landmark detection, head pose tracking, and eye gaze [17].

[3]OpenFace: a general-purpose face recognition library with mobile applications.

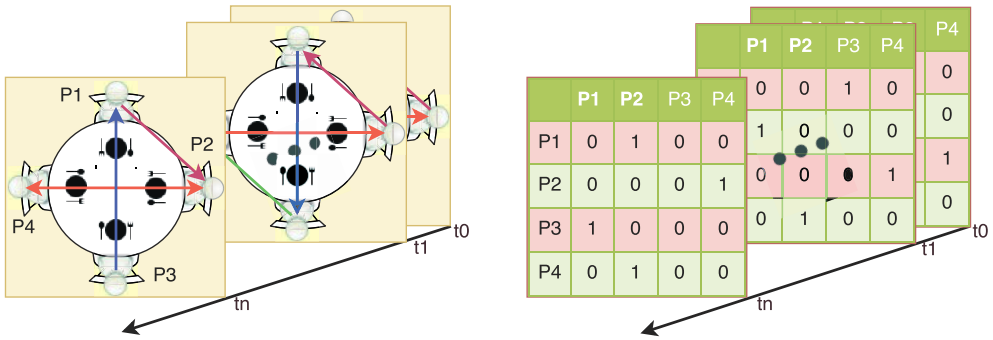


Fig. 4: Gaze estimation as an example of time variant layer. Pi is the i^{th} person; on the table, the value of (x, y) is 1 if Px is looking at Py else it is 0.

To recognize the basic emotions (happy, sad, angry, disgust, fear, and surprise), we consider the Local Binary Patterns as a feature extractor and neural network as a classifier. Finally, we adopt the OpenFace library [18] to track persons in the video.

D. Multilayer Analysis

As argued in section I, many challenges exist in face detection. Thus, we propose a multilayer source of information such as emotion, gaze, date, time, place, temperature, and social information.

Such a multilayer modeling increases the confidence of our results as it reduces the ratio of total failure; moreover, it helps for better understanding of video scenes and social contexts.

Considering the video time as a reference time entails two types of information sources. First, time-invariant source of information that does not explicitly depend on time like location, menu, date, occasion type, number of participants and their social information and relationships. Second, time-variant source information that explicitly depends on time such as gaze direction and over all emotion (see Figure 4 and 5).

1) Eye Contact (EC) Detection: As mentioned in section I, EC provides multiple functions. Even more, we can deduce many things based on the EC [12]: 1) the topic nature, in which, there is more EC in case of the topic being discussed is straightforward and less personal, whereas, there is less EC during the hesitating passages; 2) the relation between two persons, in which, there is more EC if the two persons are interested in each other.

To detect the EC between the participants, we have to identify the number of participants, denoted as n, the head $pose$, and gaze direction of each person. n is given as an external information, while the participant's head $pose$ and gaze direction are estimated using the OpenFace toolkit. After that, we need to build a square matrix with size of $n \times n$ as shown in Figure 4. In this matrix, if the values in both positions (x, y) and (y, x) equal 1, then there is an EC between participants x and y. For example, in Figure 4, EC holds between $P2$ and $P4$.

We calculate the values in the aforementioned square matrix according to the following procedure:

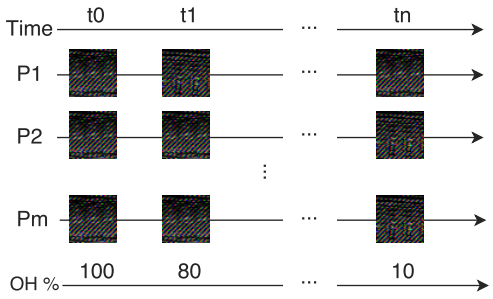


Fig. 5: Overall Emotion Estimation; Pi is the i^{th} person, OH is the overall happiness percentage.

1) Assign the reference frames as illustrated in Figure 6, where $F1$ is the reference frame of the first camera $(C1)$, $F2$ is the reference frame of the second camera $(C2)$, ${}^1F3$ is the first person $(P1)$ head $pose$ $w.r.t.$ $F1$, and ${}^2F4$ is the second person $(P2)$ head $pose$ $w.r.t.$ $F2$.
2) Compute the transformation between the frames, where ${}^1T_2$ is equal to the $pose$ of $C2$ $w.r.t.$ $F1$, ${}^1T_3$ is equal to the $pose$ of $P1$ head $w.r.t.$ $F1$, and ${}^2T_4$ is equal to the $pose$ of $P2$ head $w.r.t.$ $F2$. The transformation ${}^iT_j$ is used to transform a vector ${}^jV$ from Fj to Fi as
$$ {}^iV = {}^iT_j \times {}^jV \tag{1}$$
3) Check if Pk is staring at Pl. In particular, we have to check if the Pk gaze vector is intersecting with a sphere centered at Pl head position. Hence, both the line and the head position must be in the same reference frame. Assume that $F1$ is the reference frame, and Pk is seen by $C1$ and Pl seen by $C2$, then we transform ${}^2Vl$ to $F1$ based on equation 1 as following:
$$ {}^1Vl = {}^1T_2 \times {}^2T_4 \times {}^4Vl \tag{2}$$
Next, we model Pk head as a sphere:
$$ \|\mathbf{x} - \mathbf{c}\|^2 = r^2 \tag{3}$$
where $\mathbf{c}$ is the sphere center, r is the sphere radius, and $\mathbf{x}$ is a point on the sphere. Generically, any line can be

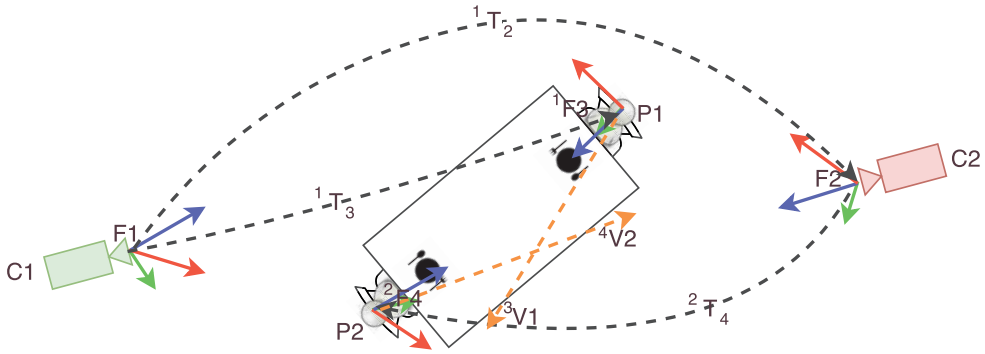
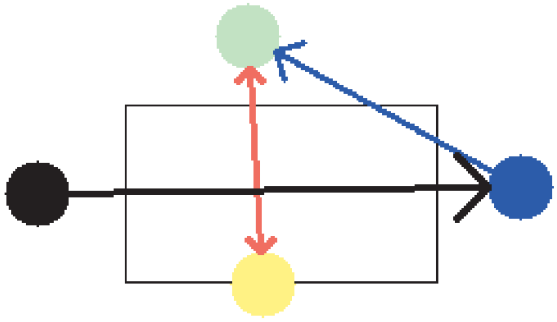


Fig. 6: Eye contact detection between two persons. $C1$, $C2$ are first and second cameras; $P1$, $P2$ are first and second persons; $F1$ is the reference frame of $C1$, $F2$ is the reference frame of $C2$; ${}^1F3$ is $P1$ head *pose* $w.r.t.$ $F1$, ${}^2F4$ is $P2$ head *pose* $w.r.t.$ $F2$; ${}^iT_j$ is the *pose* of Fj $w.r.t.$ Fi; ${}^3V1$ is the gaze direction of $P1$ $w.r.t.$ ${}^1F3$, ${}^4V2$ is the gaze direction of $P2$ $w.r.t.$ ${}^2F4$.

(a) Different Camera Views Frame.



(b) Look-At top view map.

Fig. 7: Look-at top view map built using four synchronized frames captured at time t=10s taken by four different cameras.

defined as:

$$\mathbf{x} = \mathbf{o} + d\mathbf{l} \tag{4}$$

where $\mathbf{o}$ is the origin of the line, $\mathbf{l}$ is the direction of the line, d is the distance along the line from the line starting point, and $\mathbf{x}$ is a point on the line.
Finally, we check the intersection through searching for points that are on the line and on the sphere. Thus, we combine equations 3 and 4, solve them for d, and substitute: 1) Pk head position (${}^1HPk$) as the sphere center; 2) the head position of Pl (${}^1HPl$=${}^1T_2 \times {}^2HPl$) as starting point of the line, and ${}^1Vl$ as the line direction: beginaligned

$$d = \frac{-\left({}^{\mathbf{1}}\mathbf{Vl}.\left({}^{\mathbf{1}}\mathbf{HPl} - {}^{\mathbf{1}}\mathbf{HPk}\right)\right) \pm \sqrt{w}}{\left\|{}^{\mathbf{1}}\mathbf{Vl}\right\|^2} \tag{5}$$

$$w = \begin{array}{l} \left({}^{\mathbf{1}}\mathbf{Vl}.\ {}^{\mathbf{1}}\mathbf{HPl} - {}^{\mathbf{1}}\mathbf{HPk}\right)\right)^2 \\ \quad - \left\|{}^{\mathbf{1}}\mathbf{Vl}\right\|^2 \left(\left\|{}^{\mathbf{1}}\mathbf{HPl} - {}^{\mathbf{1}}\mathbf{HPk}\right\|^2 - r^2\right) \end{array}$$

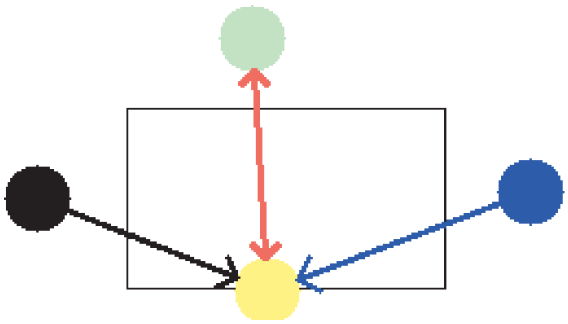
endaligned If the value of $w \in \mathcal{R}^+$, then there are two intersection points crossing the sphere and Pl is looking at Pk; otherwise the line is either tangent to the sphere or not passing through the sphere at all and Pl is not looking to Pk. We need to repeat the procedure $n(n-1)$ time to fill the squared matrix.

2) Overall Emotion Estimation: To estimate the general satisfaction of the participants, we need to evaluate the participant's overall emotion. So, we fuse various sources of information where the face recognition method, emotion recognition, and the number of participants are combined to track the participant's feeling state as illustrated in Figure 5.

E. Storing Metadata

The last step of our framework is storing both the collected external and the extracted metadata integrated with the social dimensions of the participants. This will allow us to build a video indexing and retrieval framework with rich query vocabulary so that the queries will return more semantic results.

(a) Different Camera Views Frame.



(b) Look-At top view map.

Fig. 8: Look-at top view map built using four synchronized frames captured at time t=15s taken by four different cameras.

	P1	P2	P3	P4
P1	0	0	357	0
P2	143	0	138	156
P3	519	0	0	0
P4	104	265	0	0

Fig. 9: Look-at matrix summary for the exploited video in the prototype.

III. System Prototype

As our system is in the early stage of development, we have established a small prototype acting as a proof of concept for our work. We have demonstrated the prototype using an input video recorded in a meeting room with four participants setting around a rectangle table. The input video has a duration length of 40 seconds and number of frames of 610. To precisely construct the look-at matrix, four cameras are exploited in recording a synchronized videos where the cameras are distributed on the four corners of the room and at elevation of 2.5m.

The main purpose of the prototype is to show the main strength of our system in building the eye-contact based matrix for each frame of an input video. Thus, we have leveraged the introduced eye detection method in section II. Figure 7 shows the frames of the four cameras, captured at time t=10s, with the constructed look-at top view map of the four participants. At t=10, the system has detected that the green and yellow participants look to each other, while the black one look to blue meanwhile the blue one looks to green participant. Figure 8 shows another look-at top view map configuration detected by the system at t=15s, where the green, blue, and black participants look to the yellow one. As the system provides a look-at matrix corresponding to each frame in the video, the sum of the matrix over all video frames provides a useful summary about the processed video. For instance, Figure 9 shows a look-at matrix summery computed through summing its values over 610 frames of the input video. The 357 number means how many times that the yellow participant (P1) has looked to the green participant (P3). The diagonal of the matrix is zero since the participant couldn't look to himself. The summary matrix provides useful information related to the dominate of the meeting. For instance, the yellow participant (P1) is the dominate of the meeting since the summation of the participant P1 column is the maximum.

IV. Conclusion

In this paper, we present a framework for analyzing dining events. Our framework is useful for different applications and research fields such as analyzing customer satisfaction in smart restaurants, evaluation new dining tools, and helping the sociologist in analyzing the social event videos based on the alerting functionalities like the emotion state changes, and the eye contact detection. Furthermore, having metadata repository provides more powerful and rich query vocabulary. This enhances the performance the information retrieval system as well.

As a future work, we intend to implement the framework proposed in this paper with experimenting and validating the multilayer analysis. We are planning to collect and annotate a dataset customized for our task.

References

[1] T. López-Guzmán and S. Sánchez-Cañizares, "Gastronomy, tourism and destination differentiation: a case study in spain," *Review of Economics & Finance*, vol. 1, pp. 63–72, 2012.

[2] J. Kivela and J. C. Crotts, "Tourism and gastronomy: Gastronomy's influence on how tourists experience a destination," *Journal of Hospitality & Tourism Research*, vol. 30, no. 3, pp. 354–377, 2006.

[3] M. Mizrahi, A. Golan, A. B. Mizrahi, R. Gruber, A. Z. Lachnise, and A. Zoran, "Digital gastronomy: Methods & recipes for hybrid cooking," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 2016, pp. 541–552.
[4] R. Burke and A. L. Kelly, "Molecular gastronomy," *FOOD SCIENCE AND TECHNOLOGY IRELAND*, p. 3, 2016.
[5] L. Canetti, E. Bachar, and E. M. Berry, "Food and emotion," *Behavioural processes*, vol. 60, no. 2, pp. 157–164, 2002.
[6] D. Bhattacharjee, "Adaptive polar transform and fusion for human face image processing and evaluation," *Human-centric Computing and Information Sciences*, vol. 4, no. 1, p. 4, 2014.
[7] S. Choi, J.-W. Han, and H. Cho, "Privacy-preserving h. 264 video encryption scheme," *ETRI Journal*, vol. 33, no. 6, pp. 935–944, 2011.
[8] X. Yang, G. Peng, Z. Cai, and K. Zeng, "Occluded and low resolution face detection with hierarchical deformable model," in *Computer Science and its Applications*. Springer, 2012, pp. 79–85.
[9] C. Frith, "Role of facial expressions in social interactions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3453–3458, 2009.
[10] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 2015.
[11] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 478–500, 2010.
[12] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, pp. 289–304, 1965.
[13] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
[14] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Video summarization using deep semantic features," *arXiv preprint arXiv:1609.08758*, 2016.
[15] D. Codreanu, A.-M. Manzat, and F. Sedes, "Mobile objects and sensors within a video surveillance system: Spatio-temporal model and queries," in *International Workshop on Information Management in Mobile Applications-IMMoA 2013*, 2013, pp. 52–59.
[16] J. Gao, A. G. Hauptmann, A. Bharucha, and H. D. Wactlar, "Dining activity analysis using a hidden markov model," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2. IEEE, 2004, pp. 915–918.
[17] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, March 2016, pp. 1–10.
[18] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.
[19] M. N. Asghar, F. Hussain, and R. Manton, "Video indexing: a survey," *International Journal of Computer and Information Technology*, vol. 3, no. 01, pp. 148–169, 2014.