Automatic Visual Mimicry Expression Analysis in Interpersonal Interaction

Xiaofan Sun
Human Media Interaction, University of Twente, Enschede, The Netherlands
x.f.sun@utwente.nl

Khiet P. Truong
Human Media Interaction, University of Twente, Enschede, The Netherlands
k.p.truong@utwente.nl

Anton Nijholt
Human Media Interaction, University of Twente, Enschede, The Netherlands
a.nijholt@utwente.nl

Maja Pantic
Human Media Interaction, University of Twente, Enschede, The Netherlands
Department of Computing, Imperial College London, UK
m.pantic@imperial.ac.uk

Abstract

Mimicry occurs in conversations both when people agree with each other and when they do not. However, it has been reported that there is more mimicry when people agree than when they disagree: when people want to express shared opinions and attitudes, they do so by displaying behavior that is similar to their interlocutors' behavior. In a conversation, mimicry occurs in order to gain acceptance from an interaction partner by conforming to that person's attitudes, opinions, and behavior. In this paper we describe how visual behavioral information expressed between two interlocutors can be used to detect and identify visual mimicry. We extract and encode visual features that are expected to represent mimicry in a useful way. In order to show that mimicry has indeed occurred, we calculate correlations between visual features extracted from the interactants and compare these against each other and against a baseline. We show that it is possible to visualize the occurrence of visual mimicry during the progress of a conversation which allows us to research in which situations and to what extent mimicry occurs.

1. Introduction

In psychology, mimicry research has focused on the effect of mimicry on people's social behavior and relationship judgement in social interaction. In daily social interactions, people usually automatically and continuously mimic and synchronize their behaviors with the facial expressions, voices, postures, hand gestures, and mannerisms of their interlocutors. Scientists and psychologists have long been attracted by the interesting observation that people tend to mimic the affective expressions of others. As early as 1759, Adam Smith acknowledged that people display motor mimicry because they imagine themselves in another's situation. And they even felt that such imitation was "almost a reflex". Later, such mimicry was associated with creating empathy and rapport. Theodor Lipps (1903) suggested that conscious empathy is attributable to the instinctive motor mimicry of another person's expressions of affect. Since the 1700s, researchers have collected considerable evidence that people do tend to imitate others' emotional expressions.

Recently, the automatic identification of visual mimicry has gained interest from the affective computing community, although little effort has been undertaken to develop these methods. Automatic mimicry analysis implies detection of unconscious behavior and involves the understanding of human affect and the interlocutors' relationship in a social interaction. There are several difficult tasks that need to be carried out by an automatic mimicry analyzer, for example: measurement and extraction of behavioural mimicry information, identification of an image segment as containing behavioural mimicry, and classification of mimicry expressions in labels that can be described in terms of e.g., social signal or affect categories. Combining a system that performs these operations in a multi-modal affect computing system would be a big step in the understanding the underlying or implicit human affective and social behavior.

In this paper we start with the visual channel: we measure and analyze behavioral mimicry with respect to visual cues. It may not be possible to incorporate all visual information extracted, and even so, some features may be undesirable. Hence, one of the principal and most important issues is to explore those features that are reliable and informative for representing behavioural mimicry.

The rest of this paper is organized as follows: the literature work is briefly summarized in Section 2.1. The data we use in our analyses is described in Section 2.2. The technical details about the steps of the outlined method are

first introduced in Section 2.3. In Section 2.4 the main work about obtaining regions of interest (ROI) containing complex moving is described. Then, in Sections 2.5 and 2.6 the method to extract and encode features extracted by computationally efficient representations, i.e., average motion energy (AME) and quadtree decomposition, is presented. The parameters encoded are used to analyze the occurrence of mimicry in different experiments which are described and discussed in Section 3. We conclude and give suggestions for future work in Section 4.

2. Towards visual-based mimicry detection

2.1. Related work

Mimicry is an important cue relevant to the understanding of social interaction and human affect. Hence, with respect to our ultimate goal of building an effective and robust affective computing system, developing a mimicry detection system is a primary and challenging task. In the last few decades, mimicry research has attracted researchers from different disciplines in which the human plays a central role, such as affective computing and human behavioral research. Although mimicry is a very relevant topic in these research areas, the (automatic) measurement of mimicry is still an unexplored research topic.

In practical applications, individuals are not usually required or expected to wear additional devices even if the devices are assumed to be comfortable and wearable. Hence, the focus of the current paper is on the automatic analysis of mimicry in terms of visual information. Nearly all studies on (automatic) human behavior understanding have mainly been focused on sensor-based and visual-based approaches [1], [2]. Considerable efforts have been made to develop methods analyzing human actions by visual information, here we name a few: template matching [3], [4], intensity-based feature [5], [6], shape matching [7], and spatial-temporal features [8], [9]. The approach in the current paper was initially inspired by spatial-temporal features which demonstrate that motion energy images (MEIs) can be used to incorporate temporal information into spatial images as very useful features. Bobick and Davis [8] proposed the temporal templates as models for human actions in which the idea of MEI was firstly introduced. They constructed two vector images, that is, MEI and Motion history image (MHI) by collecting a group of frames and extracting scale invariant features for encoding a variety of motion properties or characters. Usually, a MEI is a cumulative motion image which presents and emphasizes regions containing most complex and frequent motions, whereas MHI denotes current moving pixels. Also, in order to preserve temporal motion information, average motion energy (AME) is derived from MEI, by aligning and normalizing the foreground silhouettes, to depict motion in a two-dimensional space [11]. Recently, AME is widely applied in solving problems such as gender classification [12], [13] and gait classification [14], [15], [16]. Mimicry detection distinguishes itself from human action recognition in that mimicry detection focuses on recognizing similar behaviors rather than the action itself. As such, we propose to compute the behavioural similarity between actions by analyzing motion histograms for each region containing various motions in different parts of a body.

2.2. Mimicry corpus

Our proposed methods and analyses are applied to data that is drawn from a study of face-to-face discussions and conversations. 40 subjects and 3 confederates from Imperial College, London participated in this experiment. They were recruited using the Imperial College social network and were compensated 10 pounds for one hour of their participation. The role of the confederate is to elicit and show agreement and disagreement with the participant during the recording sessions. Each recording session used in this study collects data through a synchronized multimodal setup. The experiment included two sessions. In the first session, participants were asked to choose a topic from a list, which had several statements concerning that topic. Participants were then asked to write down whether they agreed or disagreed with each statement of their chosen topic before presenting their opinions (‘presentation episode’) and discussing with each other (‘discussion episode’). Each session’s duration was between 9 and 15 minutes. In the second session, the intent is to simulate a situation where participants wanted to get to know their partner a bit better and they needed to disclose personal and possibly sensitive information about themselves. Participants were given a non-task-oriented communication assignment that required self-disclosure and emotional discovery. All the durations of session 2 were between 15 and 22 minutes long.

To the best of our knowledge, this data is the first dataset aimed towards recording behavioral mimicry in multiple modalities. In each session we recorded data from the participants separately and from the two participants together, including vocal and bodily behaviors. In the visual-based channel we recorded data using 7 cameras for each person and 1 camera for both persons at the same time. The camera for both persons was used for recording an overview of the interaction, while the other 7 cameras were used for recording the two participants separately, including far-face view, near-face view, upper-body view, and whole body view with and without color. Both participants wore a lightweight and distance-fixed headset with microphone. For detecting head movements, both participants wore rigs

on their heads during recording. The rig is a lightweight, flexible metal wire frame fitted with 9 infrared LEDs. Given the face location and orientation, the nine LEDs allow us to get detailed information about the shapes of the head movements.

Three types of cameras were used in the recordings: one Allied Vision Stingray F046B, monochrome camera, with a spatial resolution of 780x580 pixels; two Prosilica GE1050C colour cameras, with spatial resolutions of 1024x1024 pixels; and 12 Prosilica GE1050 monochrome cameras, with spatial resolutions of 1024x1024 pixels. Different sets of cameras were set up to record the face regions at two distance ranges: 'Far' corresponds to a distance range for upright poses and 'Near' corresponds to forward-leaning poses. The focal length and focus of the cameras were optimized for the respective distance range. This means that the best camera view to use for a facial analysis depends on a person's body pose at any moment. The cameras were intrinsically and extrinsically calibrated.

The video sequences used in this study last on average 17.8 seconds and were all manually segmented for the presence of behavior expressions such as head movements, body postures, and hand gestures. Temporal window sizes ranging from 20 to 280 frames were all analyzed independently and the window size (i.e., the number of frames over which correlation is calculated) that suited the best segmentation was chosen. Each segmented episode includes one dyadic interaction in which one participant is the speaker and the other one is the listener. In two successive episodes, the roles of the two participants are switched. For example, in episode 1, participant 1 is speaker (S1) and participant 2 is listener (L2), then in episode 2, participant 2 must be speaker (S2) and participant 1 is speaker (L1).

2.3. Methodology

The main steps of the proposed method are illustrated in Fig.1, and can be summarized as follows: in the pre-processing phase, a set of aligned human silhouettes is extracted from the input video. Next, motion cycle extraction is performed from the input frames by calculating MEIs. Subsequently, AMEs are obtained in terms of the extracted motion period and are converted to Motion Energy Histograms (MEHs). That is, accumulated motion images (AMIs) are computed by using input image differences to represent the spatiotemporal features of the occurring motions. It is worth noting that this step is based on average motion energy computing. Then, AMIs are computed in meaningful areas which contain changes and motions instead of the whole silhouette of the human body extracted in motion cycle periods. In addition, for each AMI, quadtree decomposition is defined and performed to construct the Multi-resolution Accumulated Motion Histogram (MAMH) in order to determine regions containing most complex and frequent motions in which the pixel intensity should be the highest. The proposed values of AMIs are projected in horizontal and vertical directions. Finally, the correlations of extracted AMIs in decomposed regions are computed between two paired participants, which serve to determine the behavioural similarity for detecting mimicry in target video sequences.

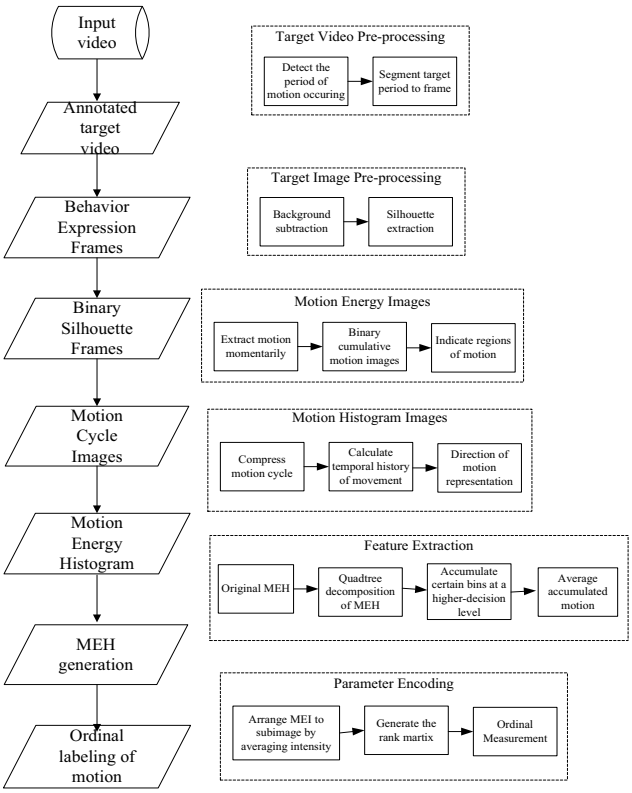


Figure 1: Outline of the proposed method

2.4. Motion representation

Motion cycle is extracted in terms of the accumulated or averaged motion energy which is only computed in areas where changes occur. Hence we propose to represent motion cycles by computing a group of AMIs. In detail, as AMI represents the time-normalized accumulative and average action energy and contains pixels with intensity values for representing motions. In the AMI, the regions containing pixels with higher intensity values denote that motions occur more frequently and with more complexity. AMI is related to MEI and MHI. However, a fundamental difference is that AMI describes the motions by using the pixel intensity directly instead of giving lower weights for older frames, or only giving equal weights for all changing areas. So compared to MHI, AMI gives us a dense and

informative representation of the occurrence of motion.

2.5. Feature extraction

In order to extract features from sub-regions containing the most complex and frequent motion rather than the whole silhouettes, we use quadtree decomposition instead of partitioning the whole body region into different parts or instead of manually localizing motion regions. Quadtree decomposition is used to divide the areas showing a lot of motion during the behavior into a large number of smaller sub-regions while areas showing little motion are divided into a small number of larger sub-regions. Areas with no motion are not divided. So quadtree decomposition can be realized as a very efficient structure, which can represent the characteristics of two dimensional values of human actions for identifying behavioural similarity.

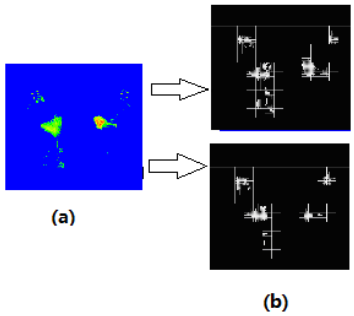


Figure 2: (a) AMEs for two participants' gestures expressed in several consecutive frames. (b) Horizontal and vertical energy histogram from top to bottom.

To identify similar behaviors between two participants in our data visually and efficiently, we use averaged motion histograms (AMHs) obtained by projecting AMI values in horizontal and vertical directions, respectively, as shown in Fig. 2. These figures, 2(a) and (b), demonstrate the identification of similarity of behaviors expressed by two participants. We did not perform spatial or temporal registration because all subjects were recorded with the same cameras in a sitting position.

2.6. Feature Encoding

After generating the quadtree decompositions and projecting AMI values in horizontal and vertical directions, we extract features from each relevant image sub-region (as informed by quadtree decomposition) in order to identify "behavior similarity". The process is repeated for all the video segmentations in the data set. Firstly, we computed the optical flow between consecutive pairs of frames of the video. Subsequently, a set of kinematic features was extracted for representing similarity between behaviors in different aspects of motion tendency present in the optical flow, such as motion intensity, vorticity, and symmetric flow fields. To this end, each sub-region was resized to an N*N sub-image by intensity averaging. The features in each sub-image were extracted and encoded to N*N matrices as shown in Fig. 3. In this figure we give an example of hand gesture mimicry and the encoding of features representing hand movements. We used a traditional template-based action recognition approach to compute behavior similarities of corresponding windows (i.e., a number of frames containing movements represented by motion cycles) in which the distances between the query action and all local windows in the target videos are computed.

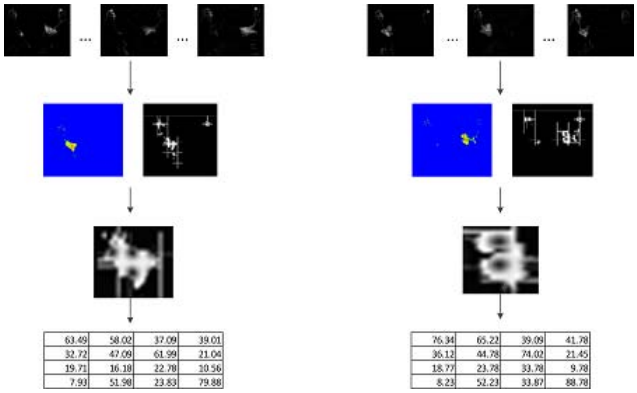


Figure 3: Two different 4 × 4 sub-images (i.e., M = 16) of the behavioral mimicry expressed by two participants in our data. The matrices show the averaged motion intensity of each sub-region.

3. Experiments

3.1. Method

We evaluate the presence of visual-based mimicry in our data by comparing pairs of interactant behaviors in each session of to each other and themselves. In this paper, we are interested in analyzing and representing the global tendency of mimicry in a face-to-face conversation. Hence, firstly, we demonstrate the changes over time in the interactants' visual behaviors and show that these become more similar to each other.

Following the comparison scheme shown in Table 1, we calculated correlations (Pearson's correlation coefficient) between the behavior patterns of the participants in different episodes and compared these correlations to each other. Firstly, we attempt to show that behavioral patterns displayed by conversational partners can be extracted and compared in order to recognize mimicry. Moreover, we demonstrate that behavioral mimicry is indeed ubiquitous in human-human conversation. In these experiments, we show that people change their postures, body movements, and gestures while interacting with others. More importantly,

these changes are made to mimic each other's visual behavior. The conversation partners' visual behaviors, represented by motion energy, optical flow, and motion histograms, were shown to converge during complete dialogue segmentation, which has been illustrated in section 3.1. We make comparisons for each individual pair of participant and confederate separately but in order to be able to make more general statements about the mimicry behavior found in our corpus, we will also look at averaged data combined from all pairs we studied.

Compare			
(A) Correlation between		(B) Correlation between	
participant in presentation	participant in discussion	participant in discussion	confederate in discussion

Table 1 Comparison scheme

3.2. Results and Discussion

One of the sub goals of our mimicry research is to identify and detect visual cues that can be used for the automatic measurement and detection of visual-based mimicry. By using this comparison scheme, we demonstrate that these paired participants gradually express more and more similar gestures and body movements with their partners in a conversation. And compared to the previous behavior expressed by the same person, we find that they indeed change their behaviors to a certain degree. Given the similarity of behavior between two paired participants, we can come to the conclusion that during a conversation, people have the tendency to adjust their own behaviors to mimic their partners' behavior. Fig. 4 shows the curves for both correlations, averaged over all pairs of participants and confederates, in which (A) the correlation between the participants' performances during the presentation and discussion episodes (solid), and (B) the correlation between the participants' performances during the discussion episode (dashed) are shown.

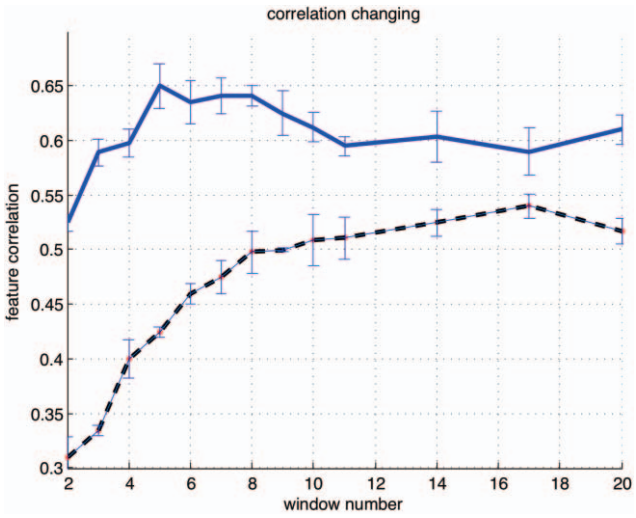
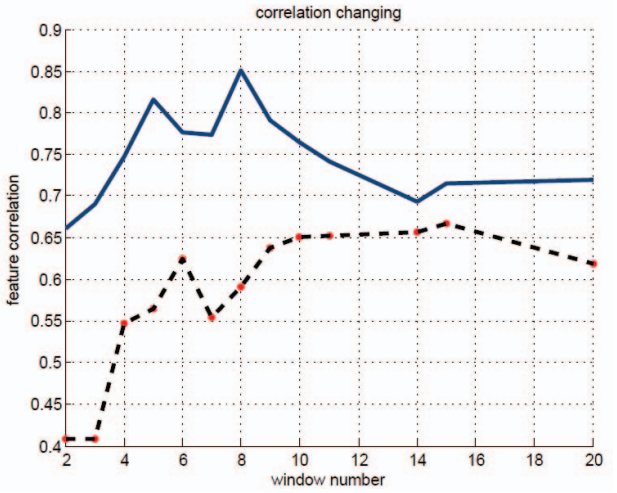
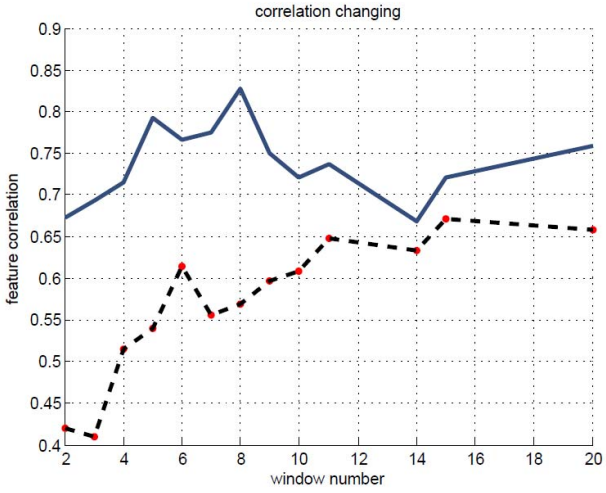


Fig. 4: Extracted visual feature correlation curves (A)=solid and (B)=dashed showing a general and global tendency to mimic each other's visual behavior – based on averaged correlations of all pairs of participants and confederates in the same session. The bars represent the standard deviations of the averaged correlations.

In addition to the correlation curves averaged over all pairs, correlation curves of two randomly chosen individual pairs of participants and confederates are also presented in Fig. 5 (a) and (b).



(a) Example of feature correlation extracted from random individual pairs of participants and confederates in the same session in our data.



(b) Example of feature correlation extracted from random individual pairs of participants and confederates in the same session in our data.

Fig. 5: (a) and (b) show extracted visual feature correlation curves of two random individual pairs of participants and confederates in the same session.

To be able to make general statements about the results obtained, we concentrated on the correlation curves that were computed and averaged over all pairs in the corpus. Firstly, correlation (A) in Fig. 4 shows mimicking behavior of the participant. The degree of visual mimicry of the participant was measured during the discussion period relative to the base line presentation period. It was found that compared to this baseline period, participants adjusted their gestures and body movements more frequently to increase similarity with the confederate's behavioural expressions when sharing similar attitudes or opinions. In addition, Fig. 4 demonstrates two apparent changing tendencies of the correlations which can be described in three phases. We observe that 1) up to window number 8, both correlations (A) and (B) are increasing, 2) between window number 8 and 17 correlation (A) is decreasing while (B) is increasing, and 3) from window number 17 on, correlation (A) is increasing while (B) is decreasing. These tendencies demonstrate that behavioral mimicry has indeed occurred to some degree. Moreover, in some specific periods, mimicry has occurred more frequently. These tendencies are consistent with existing mimicry theories (e.g., [17], [18]) and can be explained as follows. During phase 1, correlation (A) did not decrease because the participant was expected to keep similar behaviors at the beginning of the presentation and discussion. At the beginning of the discussion with the confederate, correlation (B) increased because the confederate expressed his/her opinions during discussion and while doing that made his/her behavior look more similar to that of the participant. During phase 2, correlation (B) still increased: the participant and the confederate both had a tendency to mimic each other because they had the desire to express the understanding of each other's opinion and they also had the desire to convince each other to accept their opinions in order to achieve agreement. It is common that correlation (A) shows a decrease because after a period of time in the discussion, the participant had picked up some of the behavioural characteristics of the confederate and automatically mimicked his/her behavioural expressions while expressing agreement with him/her. Finally, phase 3 occurred at the end of the discussion in which the participant and confederate were both more willing to express their own opinions because they have the desire to convince the other at the end of conversation, and unconsciously, they did so by resorting back to their own habitual behavioural styles, knowing that the end of the discussion was approaching.

4. Conclusion and future work

Our study in this paper presents a way to detect visual mimicry in a machine analysis approach. The results of our experiments show that in face-to-face interaction a confederate often adopts the gestures, postures (e.g., learning forward or backward), head movements, and mannerisms (e.g., leg crossing) of the participants with whom he/she is interacting. Specifically, we find that participants or confederates who are being mimicked are more willing to alter the way in which they interact with others to share similar affect or attitudes in order to obtain more agreement. Furthermore, those participants who had been mimicked by the confederate were more willing to present their own opinions and attitudes and perceived the interaction as more smooth. Our results are consistent with the present finding that mimicry enhances resonance, suggesting that it serves to strengthen social bonds.

In future work, for automatic visual-based mimicry detection, more kinematic-based features are needed such that analyses similar to those carried out for non-verbal vocal mimicry can be performed. We will focus on the optical flow fields in motion parts of a body, computation of kinematic features (e.g., divergence, vorticity, symmetric flow fields etc.) and the classification of these features for recognizing mimicry in order to achieve our ultimate goal to assess human affect in terms of automatic mimicry analysis. Furthermore, analyzing the effects that mimicry has on people's behavior will inform us what role adaptive mimicry plays in people's daily lives. Since feeling what others think and doing what others do is so beneficial in such a diverse array of social situations (e.g., job interviews, social networking, romantic affairs, negotiation, and selling products), it is no wonder that mimicry serves as a social

glue. Mimicry has an impact on social interaction and human affect, and hence, needs to be analyzed in order to understand these social and interactional behaviors.

Acknowledgement

We gratefully acknowledge the help of Michel Valstar, Jeroen Lichtenauer, and many others from Imperial College who helped to realize the experiments for the data collection. This work has been funded in part by the European Community's 7th Framework Programme [FP7/2007-2013] under the grant agreement no. 231287 (SSPNet). The work of Maja Pantic is also funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

References

[1] A. Briassouli and I. Kompatsiaris, "Robust temporal activity templates using higher order statistics.," IEEE Transactions on Image Processing, vol. 18, no. 12, pp. 2756–2768, 2009.

[2] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05), vol. 1, pp. 462– 469, Beijing, China, October 2005.

[3] H. J. Seo and P. Milanfar, "Detection of human actions from a single example," in Proceedings of the International Conference on Computer Vision (ICCV '09), October 2009.

[4] V. H. Chandrashekhar and K. S. Venkatesh, "Action energy images for reliable human action recognition," in Proceedings of the Asian Symposium on Information Display (ASID '06), pp. 484–487, October 2006.

[5] M. Ahmad and S.-W. Lee, "Recognizing human actions based on silhouette energy image and global motion description," in Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG '08), pp. 1–6, Amsterdam, The Netherlands, September 2008.

[6] C. Kim, "Content-based image copy detection," Signal Processing: Image Communication, vol. 18, no. 3, pp. 169–184, 2003.

[7] C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 1, pp. 127–132, 2005.

[8] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257–267, 2001.

[9] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04), vol. 3, pp. 32–36, Cambridge, UK, August 2004.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05), vol. 1, pp. 886–893, San Diego, Calif, USA, June 2005.

[11] P. S. Dhillon, S. Nowozin, and C. H. Lampert, "Combining appearance and motion for human action classification in videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09), pp. 22–29, Miami, Fla, USA, June 2009.

[12] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05), vol. 1, pp. 984–989, San Diego, Calif, USA, June 2005.

[13] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247– 2253, 2007.

[14] E. Shechtman and M. Irani, "Space-time behavior based correlation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05), vol. 1, pp. 405–412, San Diego, Calif, USA, June 2005.

[15] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07), pp. 1–8, Minneapolis, Minn, USA, June 2007.

[16] H. Ning, T. X. Han, D. B. Walther, M. Liu, and T. S. Huang, "Hierarchical space-time model enabling efficient search for human actions," IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 6, pp. 808–820, 2009.

[17] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," Journal of Personality and Social Psychology, 76(6), pp. 893-910, 1999.

[18] H. Giles, J. Coupland, and N. Coupland, Accommodation theory: communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), Contexts of accommodation. Developments in applied sociolinguistics. Cambridge: Cambridge University Press, 1991.