Facial Expression Classification Based on Local Spatiotemporal Edge and Texture Descriptors

Yulia Gizatdinova
Research Group for Emotions, Sociality and Computing, University of Tampere, Finland
yulia.gizatdinova@cs.uta.fi

Veikko Surakka
Research Group for Emotions, Sociality and Computing, University of Tampere, Finland
veikko.surakka@cs.uta.fi

Guoying Zhao
Machine Vision Group, University of Oulu, Finland
guoying.zhao@ee.oulu.fi

Erno Mäkinen
Multimodal Interaction Research Group, University of Tampere, Finland
erno.makinen@cs.uta.fi

Roope Raisamo
Multimodal Interaction Research Group, University of Tampere, Finland
roope.raisamo@cs.uta.fi

ABSTRACT
Facial expressions are emotionally, socially and otherwise meaningful reflective signals in the face. Facial expressions play a critical role in human life, providing an important channel of nonverbal communication. Automation of the entire process of expression analysis can potentially facilitate human-computer interaction, making it to resemble mechanisms of human-human communication. In this paper, we present an ongoing research that aims at development of a novel spatiotemporal approach to expression classification in video. The novelty comes from a new facial representation that is based on local spatiotemporal feature descriptors. In particular, a combined dynamic edge and texture information is used for reliable description of both appearance and motion of the expression. Support vector machines are utilized to perform a final expression classification. The planned experiments will further systematically evaluate the performance of the developed method with several databases of complex facial expressions.

Author Keywords
Human behaviour understanding, expression classification, spatiotemporal descriptor, local oriented edge, local binary pattern, facial expression, action unit, emotion.

ACM Classification Keywords
I. Computing methodologies: I.4 Image processing and computer vision; I.2 Artificial intelligence: I.2.10 Vision and scene understanding.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. For any other use, please contact the Measuring Behavior secretariat: info@measuringbehavior.org.

INTRODUCTION
During the past few decades, automatic classification or recognition of facial expressions has attracted a considerable attention in computer vision, pattern recognition and human-computer interaction (HCI). The main reason for this has been a growing demand for the development of new generation of HCI interfaces that actively perceive a user, detect visual cues of facial behaviours and use this information to initiate interaction, offer help, or assist the user in performing a certain action or task. Generally, facial expressions have been classified either as emotion-associated facial displays or action units (AU). AUs represent momentary changes in facial appearance which are brought about by single or conjoint facial muscle activations [3]. Although the research in the area of automatic expression classification has seen a lot of progress [4,11,14], there are still challenges that need to be solved in order to achieve true applicability of expression classifiers in real-world situations. Thus, there is a need for improvement of the existing classification schemes in terms of their accuracy, speed and robustness to unconstrained environmental conditions. On the other hand, there is a challenge of reliable analysis of between- and within-person variations as facial expressions vary in their appearance both across human population and within facial behaviour of a given individual. Therefore, one of the essential prerequisites for a successful development of expression recognition systems is to find descriptive, robust yet fast- and easy-to-compute facial representations.

So far, the majority of the existing research has been dealing with processing of static image data [4,11,14]. However, recent advances [13] have clearly demonstrated that humans recognize facial expressions better when the dynamics of the expression is taken into consideration. Following these findings, several spatiotemporal

approaches to expression recognition have been reported. Among others, the most frequently used approaches are hidden Markov models [1], dynamic texture descriptors [15], geometrical displacements [6] and dynamic Bayesian networks [16]. Dynamic texture descriptors combine appearance and motion features which occur at several spatiotemporal scales. Local binary pattern (LBP) texture descriptors [10] have been widely used to describe static textures and recently extended to temporal domain. LBP operator encodes various local primitives such as points, curved edges, spots, flat areas, *etc*. LBP descriptors have been used for facial representation and reported to provide a number of important advantages like local processing, multi-scale representation, robustness to monotonic grey-scale changes and simple computation [15]. LBP descriptors, which are calculated on three orthogonal planes (TOP) of the image sequence, result in LBP-TOP dynamic texture descriptors and have been used to recognize a limited set of posted emotion-associated facial expressions [15]. The reported classification performance has approached the classification performance of a human observer with more than 90% of classification success. Recently, several attempts have been done in order to enhance LBP-TOP operator by applying it to gradient [9] or Gabor-decomposed [7,8] images.

In this paper, we present an ongoing research that aims at development of a novel spatiotemporal approach to facial expression classification in video. The novelty comes from a new facial representation that is based on local spatiotemporal feature descriptors. In particular, a combined dynamic edge and texture information is used for interpretation of the appearance and dynamics of facial behaviour. Differently from previous studies, we aim to enhance LBP-TOP descriptor by combining it with a dynamic edge descriptor based on local oriented edges (LOE) [5]. LOE descriptors encode local edges of multiple orientations at different resolution levels. In the earlier studies [5], LOE descriptors have been successfully applied to the task of facial feature localization from images with complex facial expressions. Therefore, dynamic LOE-TOP descriptors are expected to improve the descriptive property of LBP-TOP operator.

In the proposed method, LOE- and LBP-TOP feature vectors (histograms) are calculated in spatiotemporal domain and concatenated into a single histogram. The received LOE-LBP-TOP spatiotemporal histogram is further used for enhanced yet relatively compact and easy-to-compute representation of the appearance and motion of the expression in video data. This idea differs from previous works in which edge map of the image is constructed first and after that encoded by LBP-TOP operator. In the latter case, the parameters which are used to construct edge map of the image (*e.g.* filter size, degree of image smoothing, *etc*.) ultimately define the amount of information available to LBP operator. If parameters are selected too globally, local features which are important for facial expression recognition (*e.g.* small wrinkles, protrusions, shadings, *etc*.) [3] may get lost. That is why we see a promising way to calculate LOE- and LBP-TOP histograms both on raw intensity images and concatenate them at the later stages of processing. This preserves the possibility for LOE operator to use somewhat more global information while ensuring that LBP operator captures all local statistics. In the remaining of the paper, we introduce a methodology for classification of facial expressions in video. The classifier is based on the proposed local spatiotemporal edge and texture facial representation. Support vector machines are used to classify an expression into one of the expression categories. At the end of the paper we discuss shortly about the planned testing of the new method.

METHODOLOGY

The entire methodology of the proposed expression classification is shown in Figure 1. The video is considered as a sequence of images denoted by $\{I(x, y, t)\}_i$, where the first two dimensions (x, y) represent spatial domain and the temporal dimension t is defined by the number of images (frames) i in the sequence. As the first step, facial area is detected using Viola-Jones real-time face detector [12] that has proved to work robustly with detection accuracy suitable for a realistic application. After that 3D image grid is constructed that divides the entire image sequence into a number of video volumes.

LOE- and LBP-TOP spatiotemporal descriptors are then calculated on TOP of each video volume. In Figure 1, the orthogonal planes are depicted as central slices of a video volume and are defined as XY (violet), XT (red) and YT

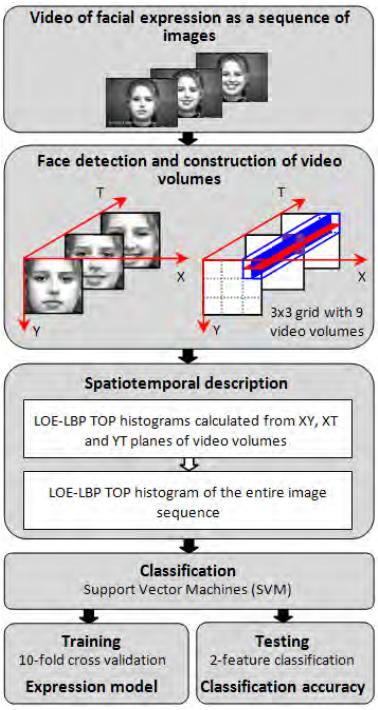

Figure 1. Block-scheme of the proposed method.

(blue). From these planes, LOE-LBP-XY, LOE-LBP-XT and LOE-LBP-YT histograms are calculated. In this process, LBP histograms are added to the end of LOE histograms (or vice versa) for every plane in each video volume. The histograms obtained from three planes compose a feature vector in a form of concatenated LOE-LBP-TOP volume histogram. After this, the concatenated LOE-LBP-TOP volume histograms are combined into a single concatenated histogram of the entire image sequence that represents a spatiotemporal signature of a given expression. Video volumes can have a certain degree of overlap. As noted in [15], the use of volume overlaps leads to the improved performance of the expression classifier.

A non-linear SVM classifier [2] with radial basis function (RBF) kernel is used to classify a given expression into one of the expression categories. In the training phase, short video clips of facial expressions are used to construct expression models for each expression category. Training of the classifier happens with concatenated LOE-LBP-TOP histograms obtained from video data as described above. The best parameters C and γ of the classifier are selected empirically by the 10-fold cross validation procedure in a grid approach. In the testing phase, two-feature classification (one-against-one) scheme is used for multi-feature classification. The idea is to create classifiers for every pair of two classes and the aim is to learn more specific and discriminative features from each pair. This way, *N*-feature classification problem is decomposed to *N(N-1)/2* two-feature problems and a voting scheme is used to define the result of the "winning" classifier.

LBP-TOP Spatiotemporal Descriptors

Local binary pattern (LBP) operator [10] is a grey-scale texture measure that is derived from the image by thresholding pixel values in a local neighbourhood of arbitrary circular shape. *LBP(P,R)* operator produces 2^P different binary codes that are formed by *P* pixels in the local neighbourhood of radius *R*. The derived binary numbers encode local texture primitives such as points, curved edges, spots, flat areas, *etc*. After computing LBP codes for the whole image, a histogram is constructed that describes occurrences of binary codes in the image. This way, LBP histogram represents distribution of local texture patterns over the whole image. For computation of spatially enhanced LBP histogram, binary codes are calculated in separate blocks of 2D image grid of size $N \times M$. The resulted $N \cdot M$ block histograms are further concatenated into a single spatially enhanced LBP histogram that holds information about occurrences of local texture patterns in different parts of the image.

For computation of LBP-TOP histogram [15], statistics on three planes (XY, XT and YT) are computed and then concatenated into a single histogram. Correspondently, the resulting feature vector (histogram) is of $3 \cdot 2^P$ length. Figure 2 illustrates the construction of LBP-TOP histogram. In such a scheme, LBP descriptors encode the appearance and

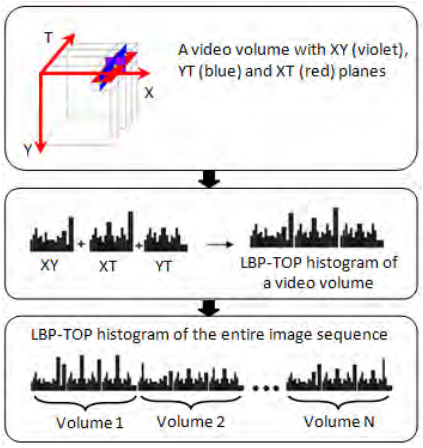

Figure 2. Spatially-enhanced LBP-TOP histogram.

motion of the expression in three directions, incorporating spatial information in LBP-XY histogram and spatiotemporal co-occurrence statistics in LBP-XT and LBP-YT histograms. In our implementation, LBP-TOP histograms are computed in each volume of the input video, resulting into spatially enhanced LBP-TOP histogram of the image sequence as shown in Figure 2. LBP-TOP operator is expressed as:

$$LBP-TOP(P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T) \quad (1)$$

where the notation (PXY, PXT, PYT, RX, RY, RT) denotes a neighbourhood of *P* points equally sampled on a circle of radius *R* on XY, XT and YT planes respectively. The best results for emotion-associated expression classification have been obtained [15] with uniform LBP-TOP (8,8,8,3,3,3) operator on 9×8 grid with 70% of volume overlap.

LOE-TOP Spatiotemporal Descriptors

Local oriented edge (LOE) operator [5] is used to detect local edges by convolving pixel values in a local neighbourhood with a set of convolution kernels. Convolution kernels result from differences of two Gaussians with shifted centres and encode the orientation of a local edge in the central pixel of the neighbourhood. Figure 3 shows the orientation template that defines 16 different edge orientations with a step of 22.5°. Before computing LOE descriptors, image is smoothed with a Gaussian filter in order to eliminate noise. LOE operator

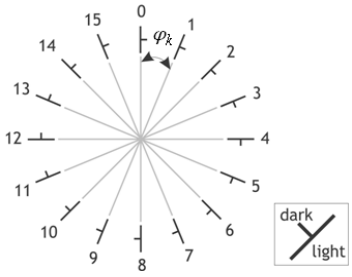

Figure 3. Orientation template, $\varphi_k = k \cdot 22.5°$, $k = 0 \div 15$.

Proceedings of Measuring Behavior 2010 (Eindhoven, The Netherlands, August 24-27, 2010)
Eds. A.J. Spink, F. Grieco, O.E. Krips, L.W.S. Loijens, L.P.J.J. Noldus, and P.H. Zimmerman

outputs a histogram that describes occurrences of local edges of certain orientation in the image.

The idea of LOE histogram calculation in the spatiotemporal domain is similar to the spatial case described above. A reduction of dimensionality of LOE-TOP histogram is done similarly to LBP-TOP. Thus, LOE descriptors are computed on the orthogonal planes XY, XT and YT of each video volume, concatenated first into volume histograms and, finally, into spatially enhanced LOE-TOP histogram of the entire image sequence. LOE-TOP operator is denoted as:

$$LOE-TOP(\varphi_k, \sigma, N, l, P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T) \quad (2)$$

where φ_k is angle of the Gaussian rotation, $\varphi_k = k \cdot 22.5°$, $k = 0 \div 16$; σ is a root mean square deviation of the Gaussian distribution; *N* denote a size of the filter; *l* defines a level of image smoothing (resolution). The parameter values for facial feature detection from expressive images have been reported [5] as $\sigma = 1.2$, $N = 7$ and $l = 2$.

DISCUSSION AND FUTURE WORK

A novel spatiotemporal approach to classification of facial expressions in image sequences is presented. The method is based on the combined LOE-LBP-TOP local spatiotemporal representation. The main advantage of the new representation is its capacity of modelling local dynamic statistics that originates from LOE and LPB descriptors. The proposed LOE-LBP-TOP representation reflects more local variations as compared to original LBP-TOP approach and, therefore, is expected to achieve better performance. Based on the theoretical methodology presented, it is reasonable to expect the combined LOE-LBP-TOP descriptors to constitute a promising facial representation for expression classification purposes.

In the future, we plan to apply the developed methodology to recognize an extensive set of emotion-associated and AU-coded expressions. The results of a systematical testing of the new method will be compared against existing studies that applied LBP-TOP on, for example, gradient and Gabor-decomposed images.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (projects 129354 and 129502) and the University of Tampere.

REFERENCES

1. Aleksic, P.S. and Katsaggelos, A.K. Automatic facial expression recognition using facial animation parameters and multi-stream HMMs. *IEEE Trans. Information Forensics and Security 1*, 1 (2005), 3.
2. Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines (2001). http://www.csie.ntu.edu.tw/~cjlin/libsvm.
3. Ekman, P. and Friesen, W. *Facial Action Coding System (FACS): A technique for the measurement of facial action*. Consulting Psychologists Press, Palo Alto, California, 1978.
4. Fasel, B. and Luettin, J. Automatic facial expression analysis: A survey. *J. Pattern Recognition 36*, 1 (2003), 259-275.
5. Gizatdinova, Y. and Surakka, V. Feature-based detection of facial landmarks from neutral and expressive facial images. *IEEE Trans. Pattern Analysis and Machine Intelligence 28*, 1 (2006), 135-139.
6. Kotsia, I. and Pitas, I. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. Image Processing 16*, 1 (2007), 172-187.
7. Lei, Z., Liao, S., He, R., Pietikäinen, M. and Li, SZ. Gabor volume based local binary pattern for face representation and recognition. In *Proc. Int. Conf. Automatic Face and Gesture Recognition* (2008), 1-6.
8. Liao, S., Fan, W., Chung, A.C.S., and Yeung, D.-Y. Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features. In *Proc. IEEE Int. Conf. Image Processing* (2006), 665-668.
9. Mattivi, R. and Shao, L. Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor. *Lecture Notes in Computer Science 5702*. In *Proc. Int. Conf. Computer Analysis of Images and Patterns* (2009), 740-747.
10. Ojala, T., Pietikäinen, M. and Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence 24*, 7 (2002), 971–987.
11. Pantic, M., Rothkrantz, J. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Analysis and Machine Intelligence 22*, 12 (2000), 1424–1445.
12. Viola, P. and Jones, M. Robust real-time face detection. *Int. J. Computer Vision 57*, 2 (2004), 137–154.
13. Wehrle, Th., Kaiser, S., Schmidt, S., Scherer Kl.R. Studying the dynamics of emotional expression using synthesized facial muscle movements. *J. Personality and Social Psychology 78*, 1 (2000), 105-119.
14. Zeng, Z., Pantic, M., Roisman, G.I. and Huang T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence 31*, 1 (2009), 39–58.
15. Zhao, G. and Pietikäinen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence 29*, 6 (2007), 915–928.
16. Zhang, Y. and Ji, Q. Active and dynamic information fusion for facial expression understanding from image sequence. *IEEE Trans. Pattern Analysis and Machine Intelligence 27*, 5 (2005), 699–714.