Hough Forest-Based Facial Expression Recognition from Video Sequences

Gabriele Fanelli, Angela Yao, Pierre-Luc Noel, Juergen Gall, and Luc Van Gool

BIWI, ETH Zurich
VISICS, K.U. Leuven
{gfanelli,yaoa,gall,vangool}@vision.ee.ethz.ch,
noelp@student.ethz.ch
http://www.vision.ee.ethz.ch,
http://www.esat.kuleuven.be/psi/visics

Abstract. Automatic recognition of facial expression is a necessary step toward the design of more natural human-computer interaction systems. This work presents a user-independent approach for the recognition of facial expressions from image sequences. The faces are normalized in scale and rotation based on the eye centers' locations into tracks from which we extract features representing shape and motion. Classification and localization of the center of the expression in the video sequences are performed using a Hough transform voting method based on randomized forests. We tested our approach on two publicly available databases and achieved encouraging results comparable to the state of the art.

Keywords: Facial expression recognition, generalised Hough transform.

1 Introduction

Computers, already part of our lives, will never seamlessly blend in until they are able to communicate with people the same way as we do. This means that machines should be able to sense and reproduce affective behavior, i.e., understand the user's feelings and react accordingly.

Facial expressions represent one of the most important ways for humans to transmit and recognize feelings and intentions. Since the seminal work of Darwin [1], the field has fascinated psychologists, neuroscientists, and lately also computer scientists. Paul Ekman's studies in the 1970's [2] suggested that all emotions belong to a rather small set of categories. These "basic" emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*) are expressed by the same facial movements across different cultures, and therefore represent an appealing choice when designing automatic methods for facial expression classification.

The ability for a computer system to sense the user's emotions opens a wide range of applications in different research areas, including security, law enforcement, medicine, education, and telecommunications [3]. However, it is important not to confuse human emotion recognition from facial expression recognition: the latter is merely a classification of facial deformations into a set of abstract classes, solely based on visual information. Indeed, human emotions can only be inferred from context, self-report, physiological indicators, and expressive behavior which may or may not include facial expressions [4].

K.N. Kutulakos (Ed.): ECCV 2010 Workshops, Part I, LNCS 6553, pp. 195–206, 2012.
© Springer-Verlag Berlin Heidelberg 2012

There are two main methodological approaches to the automatic analysis of facial expressions [5]: *Judgment-based approaches* attempt to directly map visual inputs such as images or video sequences into one of a set of categories, while *sign-based approaches* describe facial expressions by means of coded facial actions, e.g., methods based on Ekman's Facial Action Coding System [6], which represent face deformations by activations of a set of Action Units corresponding to facial muscles movements.

This paper presents a judgment-based method for the classification of facial expressions into one of the basic emotion labels. Having seen the success of Hough transform-based methods for object detection [7–11] and action recognition [12], we investigate a Hough transform voting approach applied to the task of facial expression recognition. After having localized and normalized the faces with respect to the eyes' centers, the image sequences are arranged into cuboids, or, extending the notation of [12], *expression tracks*. These are a representation of the face which is invariant to location, scale, and (in-plane) rotation. On the tracks, classification is performed by casting votes for the expression label and temporal center of the expression.

To our knowledge, this is the first time that a Hough-voting approach is applied to the task of facial expression recognition. As in [12], the voting is performed by a forest of random trees, or Hough forest [9], and a mapping is learnt between densely sampled spatio-temporal features and the center of the expression in the video sequence. The trees are trained in a multi-class fashion and can therefore discriminate between different classes simultaneously. The leaf nodes can vote for each class and represent a discriminative codebook sharing features across classes.

Compared to the task of action recognition from video [12], facial expressions (even when posed) present more subtle differences and are therefore more difficult to classify. Additions to [12] include the normalization of the tracks with respect to rotation and the use of more discriminative shape features. In the experiment section, we thoroughly evaluate our system on standard databases of facial expressions. Our results are comparable to state-of-the-art methods, which supports our idea that Hough-voting approaches are promising tools for advancing in the field of automatic facial expression recognition.

2 Related Work

Suwa et al. [13] were the first to attempt at automatically recognizing facial expressions in 1978. Since then, the new field of research has seen a steady growth, gaining momentum in the 1990's thanks to the advances in algorithms for face detection and the availability of cheaper computing power, as the surveys of [5] and [14] show.

The initial face localization and normalization step, common to virtually all approaches to facial expression recognition from video, serves to achieve a representation of the face invariant to scale, translation, in-plane rotation, and illumination conditions. The literature is rich with approaches which normalize the images based on the location of the face [15], of the eyes [16], or thanks to facial features tracking methods [17, 18]. After the normalization stage, the remainder of an automatic facial expression recognizer consists of feature extraction and classifier design. Features need to minimize variation within the expression classes while maximizing the variation between different classes. Features can be computed from geometric measurements, e.g., from the locations of specific points tracked on the face throughout the sequence [17, 19]. Alternatively, image-based features can be extracted from texture patches covering either the whole face (holistic) or specific subregions of it (local). Commonly employed feature extraction methods from facial textures and their temporal variations include optical flow [20, 21], Gabor filter responses [16, 22], and Linear Binary Patterns [23, 24]. For the actual classification, AdaBoost and its combination with Support Vector Machines have recently gained a lot of attention [16, 25]. Other popular approaches include nearest-neighbor searches [15] and Hidden Markov Models [17, 19, 24].

Trees and forests have been previously used for action recognition, but only as indexing structures for performing efficient nearest-neighbor searches [26, 27]. Following [12], we build a holistic, image-based method for recognizing facial expressions which uses a random forest to learn the mapping between 3D video patches and votes in a Hough space for the label and the temporal location of the expression.

3 Voting Framework for Facial Expression Recognition

Having seen the successful application of random forests and Hough voting to action recognition [12], we investigate its performance on the task of recognizing facial expressions. In order to introduce the basics of the method, we assume our data to be already arranged into expression tracks, i.e., the face images are cropped and aligned as shown in Fig. 1*(a)*. Section 4 provides insights on how this normalization is performed.

3.1 Training

We start from the assumption of having a set of training expression tracks available for each class $c \in C$. Training sequences are annotated for the expression label and the temporal location of the apex in the track. In order to learn the mapping between patches from the expression tracks and a voting space, we use the Hough forest method [9]. Previously developed for 2D single-class object detection, Hough forests have recently been extended to handle multi-class detection in the spatio-temporal domain and applied to the task of action recognition [12].

Randomized Hough forests are composed of a set of random trees. A tree T is constructed from a set of patches $\{\mathcal{P}_i = (\mathcal{I}_i, c_i, \boldsymbol{d}_i)\}$ randomly sampled from the training sequences. $\mathcal{P}_i$ is a 3D patch (e.g. of $20 \times 20 \times 3$ pixels) sampled from the expression track as illustrated by the colored cuboids in Fig. 1. $\mathcal{I}_i$ are the multi-channel features extracted at a patch, i.e., $\mathcal{I}_i = \left(I_i^1, I_i^2, ..., I_i^F\right) \in \mathbb{R}^4$, where each I_i^f is feature channel f at patch i and F is the total number of feature channels. c_i is the expression label ($c_i \in C$) and $\boldsymbol{d}_i$ is a 3D displacement vector from the patch center to the center of the expression in the sequence. Figure 1 shows an expression track *(a)* and sample 3D patches extracted from it *(b)*, voting for both the expression class and the center of the expression in the sequence.

During training, the trees are built recursively starting from the root, as in the standard random forest framework [28]. Each non-leaf node is assigned a binary test based on the patch appearance $\mathcal{I}$; depending on the test's result, the training patches are split into the children nodes. The process is iterated until a leaf is created, either from reaching a maximum tree depth or from reaching a minimum number of remaining patches.

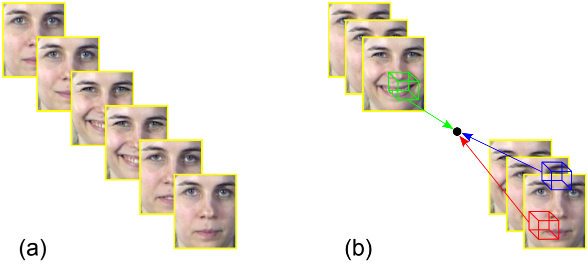


Fig. 1. Hough voting in the case of expression recognition. *(a)* Sample facial expression track. *(b)* Sample 3D patches drawn from the track, voting for the expression label and its spatio-temporal center.

As tests, we use simple comparisons of two pixels at locations $\boldsymbol{p} \in \mathbb{R}^3$ and $\boldsymbol{q} \in \mathbb{R}^3$ in feature channel f with some offset τ. For node B, the corresponding test t_B is defined as:

$$t_{B,f,\boldsymbol{p},\boldsymbol{q},\tau}(\mathcal{I}) = \begin{cases} 0 & \text{if } I^f(\boldsymbol{p}) < I^f(\boldsymbol{q}) + \tau \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Similar to [12], each binary test is assigned in order to either optimize class-label or center offset uncertainty. To this end, a set of binary tests $\{t^k\}$ is generated at each node, with random values for f, $\boldsymbol{p}$, $\boldsymbol{q}$ and τ, and evaluated on all the patches arriving at that node. The optimal test (the minimizing class label or center offset uncertainty in the split of the patches) is then chosen and assigned to the node.

When the training process is over, the leaves will store ${p_c}^L$ (the proportion of patches per class label which reached the leaf) and ${D_c}^L$ (the patches' respective displacement vectors). Patches extracted from different classes arriving to the same leaf share the same features. The proportion of patches per class label at a leaf note can be used as class probabilities ${p_c}^L$ which can indicate the degree of sharing among classes.

3.2 Facial Expression Classification

At classification time, patches are densely extracted from the test track and sent through all trees in the forest. The patches are split according to the binary tests in the non-leaf nodes and, depending on the reached leaf, cast votes proportional to p_c for the expression label and votes for the spatio-temporal center of each class c according to a 3D Gaussian Parzen window estimate of the center set vectors D_c. Votes from all patches are integrated into a 4D Hough accumulator, exemplified in the left part of Figure 2 for a sequence expressing anger. The dark spots correspond to the probabilistic votes that have been cast by the patches and accumulated in the four-dimensional space (x and y location, time, and class label). As the track has already been localized in space, we marginalize the votes into a 2D accumulator for only class label and time. The local maximum in the remaining Hough accumulator finally leads to the classification prediction, as displayed in Fig. 2, right. For a formal description of the voting process, we refer the reader to [12].

Time-scale invariance could be achieved by up-sampling or down-sampling the tracks, and then applying the same Hough forest to label expressions displayed at different speeds. However, the system has some tolerance built in through the variation in speed observed in the training data and we therefore did not consider multiple time scales.

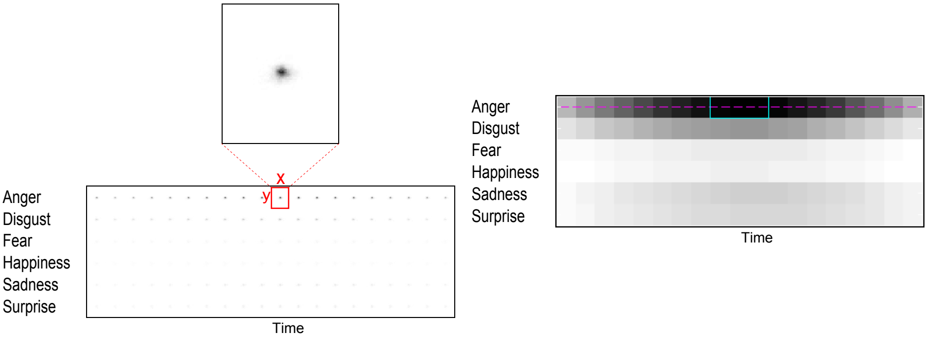


Fig. 2. *Left:* An example of a 4D Hough image, output of the voting for a clip displaying anger. The dark dots represent clusters of votes. *Right:* Example Hough voting space reduced to the two dimensions expression class and time. The maximum (in dark) is taken as the expression label and temporal location.

4 Building the Expression Tracks

In order to arrange the data in the required normalized expression tracks, we align the faces based on the locations of the eyes. Face are rotated and scaled so that the eyes lie on the same horizontal line and have the same inter-ocular distance. The invariance to rotation, an addition to the work of [12], is necessary for the task of expression recognition, which are more subtle and harder to recognize than human actions. When ground-truth annotation of the eye locations is not available, we employ a completely automatic method, i.e., the first part of the system described in [11]: after tracking the face by means of an online-boosting method [29], the eyes are localized thanks to their unique shape [30] and tracked using a pair of Kalman filters. The automatic procedure is shown in the left part of Figure 3.

4.1 Feature Extraction

For classification, simple features such as color, greyscale intensity, spatial gradients along the x and y axis, and frame to frame optical flow, were used in [12]. In our approach, inspired by the work of Schindler [31], we extract features separately representing the form and the motion of the face in the expression track. The information about form comes from the responses of a bank of log-Gabor filters. In comparison to standard (linear) filters, log-Gabor filters show an improved spectrum coverage with fewer scales [32]. The response g at position (x,y) and spatial frequency w is:

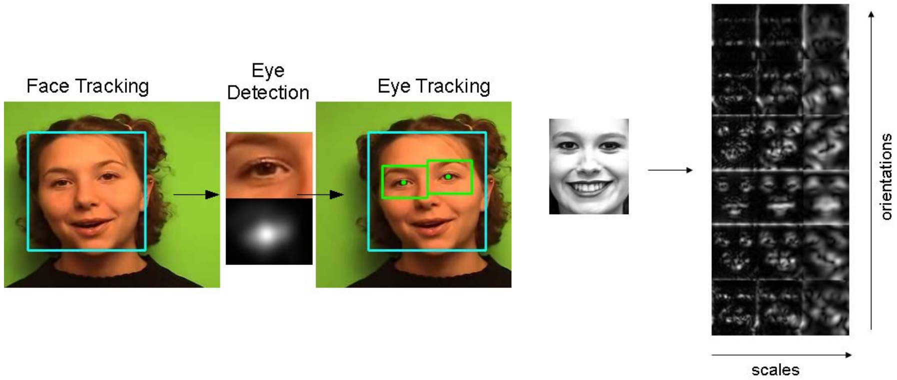


Fig. 3. *Left:* automatic face and eye tracking employed for the normalization of the facial images. *Right:* example log-Gabor responses extracted from a normalized expressive face.

$$g^{w}(x,y) = \frac{1}{\mu} e^{-\frac{\log(w(x,y)/\mu)}{2\log\sigma}} \quad , \tag{2}$$

where μ is the preferred frequency and σ a constant used to achieve an even coverage of the spectrum. We use a bank with 3 scales ($\mu \in \{2, 4, 8\}$ pixels) and 6 equally spaced orientations ($\phi \in \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ\}$), keeping only the response's magnitude $\|g^{w}(x,y)\|$ as descriptor. Example responses of the filters applied to one frame of an expression track are shown in the right part of Figure 3.

For the information regarding motion, dense optic flow is computed at every frame by template matching, using the L_1-norm, considering 4 directions. Assuming that our expression tracks always start with a neutral face, we compute the optical flow both with respect to the previous frame (frame2frame) and to the first frame of the track (frame2first).

In order to increase robustness to translation and to reduce the dimensionality of the feature space, both the shape and motion feature images are down-sampled by max-pooling, also known as winner-takes-all [33]:

$$h(x,y) = \max_{(i,j) \in \mathcal{G}(x,y)} \left[\, g(i,j) \,\right] \, , \tag{3}$$

where $\mathcal{G}(x,y)$ denotes the neighborhood of pixel (x,y). We use a window of size (3×3).

5 Experiments

We trained and tested our facial expression recognition system on the Cohn-Kanade database [34] and the MMI database [35]. Both datasets contain videos of posed facial expressions, with subjects facing the camera and under controlled lighting conditions.

The Cohn-Kanade database consists of greyscale video sequences of 100 university students, 65% of which were female. The videos always start with a neutral face and end at the apex, i.e., the maximum intensity of the expression. For our study, we selected sequences which can be labeled as one of the basic emotions and which are longer than 13 frames, for a total of 344 videos depicting 97 subjects, each performing 1 to 6 facial expressions.

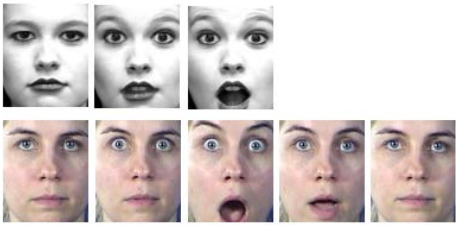


Fig. 4. Sample frames extracted from sequences depicting surprise in the Cohn-Kanade database (top) and MMI database (bottom). Note how the MMI database contains not only the transition from the neutral face to the apex of the expression, but also the offset leading back to the neutral state at end of the sequence.

The MMI database [35] is a constantly growing, web-searchable set of color videos containing both posed and spontaneous emotions. We selected the subset of (posed) videos labeled as one of the six basic emotions, while discarding all others labeled only in terms of Action Units. The resulting set is comprised of 176 videos of 29 people displaying 1 to 6 expressions. The subjects differ in sex, age, and ethnic background; moreover, facial hair and glasses are sometimes present. The main difference between the MMI and Cohn-Kanade databases is that the MMI sequences do not end at the expression's apex, but return to a neutral face. An example sequence from both dataset is shown in Figure 4, with the Cohn-Kanade at the top and MMI database at the bottom.

As explained in section 4, both databases have been aligned to the eye center locations. For the Cohn-Kanade database, ground truth manual annotations are provided by [36], while no such labeling is available for the MMI database, on which we use the eye tracking method of [11]. In both cases, the facial images are normalized to an inter-ocular distance of 25 pixels, resulting in 45×55 pixels images. Expression tracks need to be labeled with both spatial and temporal center of the expression. The center in the image plane is assumed to correspond to the center of the face. The temporal center should ideally be located at the expression apex, therefore we take the last frame for the Cohn-Kanade database and the middle frame in the case of the MMI database. We train and test on all frames from the Cohn-Kanade dataset, which has an average sequence length of 18 frames, while selecting only 20 frames in the middle of each sequence for the MMI database, which has an average length of 79 frames.

For all of the following experiments, we performed subject-independent 5-fold cross validations, i.e., making sure that the same subjects did not occur in both training and test sets, and present here the results averaged over all five iterations. Forests always contained only 5 trees; indeed, adding more trees improved the results only slightly.

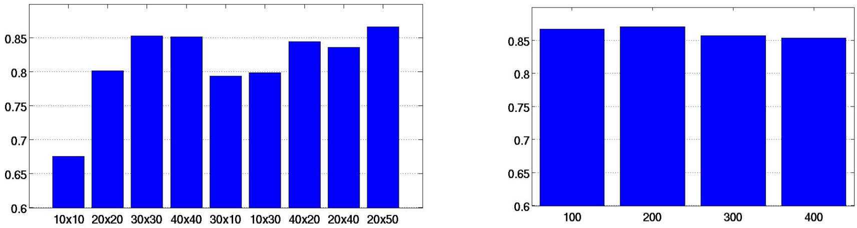
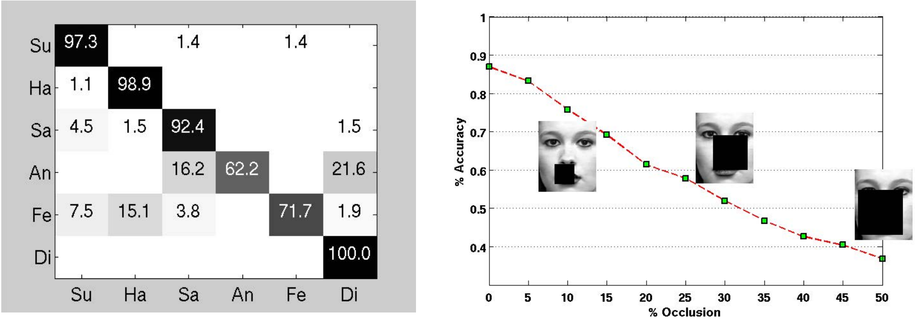


Fig. 5. *Left:* Influence of the patch size on the overall recognition rate. Larger, rectangular patches, give the best results. *Right:* Recognition accuracy as a function of the number of (20 × 50 × 2) patches sampled from each sequence during training.



	Su	Ha	Sa	An	Fe	Di
Su	97.3		1.4		1.4	
Ha	1.1	98.9				
Sa	4.5	1.5	92.4			1.5
An			16.2	62.2		21.6
Fe	7.5	15.1	3.8		71.7	1.9
Di						100.0

Fig. 6. *Left:* Confusion matrix for the Cohn-Kanade database. Expressions such as disgust and surprise are well recognized, while most of the confusion arises from the anger/disgust and fear/happiness classes. *Right:* Recognition rate for the Cohn-Kanade database, as a function of the percentage of occlusion.

Among the parameters of our proposed method are the size and shape of the patches. We ran some experiments varying the patches' spatial size and shape, while keeping the number of patches fixed to 100 and the temporal dimension to 2 frames. In Figure 5, left, the bars represent the recognition rate as a function of the size and shape of the sampled patches, as achieved on the Cohn-Kanade database. As can be noted, larger patches produce better results than smaller ones and rectangular shapes outperform squared ones. The best results (86.7%) are achieved with 20 × 50 patches, i.e., vertical rectangles covering almost half of the face.

Increasing the number of training patches per sequence did not influence much the recognition accuracy. Figure 5, right, shows that the accuracy increases only when moving from 100 to 200 patches, while it actually slightly decreases when more patches are used. We also tested the influence of the temporal length of the patches, but did not experience significant changes in the expression recognition accuracy. All results shown in the rest of the section are achieved by sampling 200 patches of size 20 × 50 × 2.

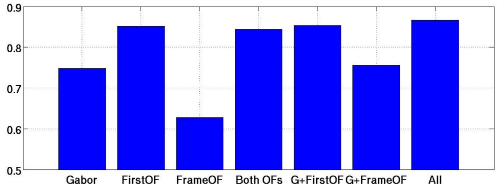
Figure 6 left shows the confusion matrix obtained by our method when applied to the Cohn-Kanade dataset. On average, we recognize the correct expression 87.1% of

Table 1. The results of our method are comparable with other works on automatic expression recognition. The accuracy is given for each expression class separately and on average.

	Our approach	Yeasin[21]	Buenaposada[15]	Aleksic[17]
SURPRISE	97.3%	**100.0%**	**100.0%**	**100.0%**
HAPPINESS	**98.9 %**	96.6%	98.8%	98.4%
SADNESS	92.4%	**96.2%**	82.0%	**96.2%**
ANGER	62.2%	**100.0%**	78.4%	70.6%
FEAR	71.7%	76.4%	73.9%	**88.2%**
DISGUST	**100.0 %**	62.5%	87.9%	97.3%
AVERAGE	87.1%	90.9%	89.1%	**93.6%**

the time; in particular, disgust is always correctly recognized. Fear and anger are the most confused labels, and are mainly mistaken for happiness, respectively disgust.

To assess the robustness of the method to partial occlusions, we removed (set to zero) the information in each feature channel falling under a cuboid. The cuboids are as long as the sequences, and cover a specific percentage of the image plane. For each sequence, the cuboid location on the 2D image plane was randomly chosen. We ran 5 trials for each percentage of occlusion, and present the averaged results in Figure 6, right. It can be noted how the performace decreases slowly as the occlusion becomes greater. At 15% occlusion, the accuracy is still around 70%, falling below 50% only when more than 30% of the face is removed. Sample frames help visualizing the amount of occlusion introduced.



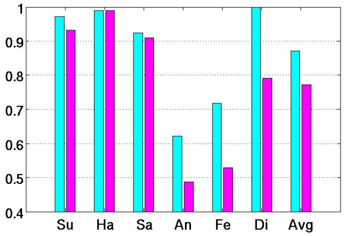
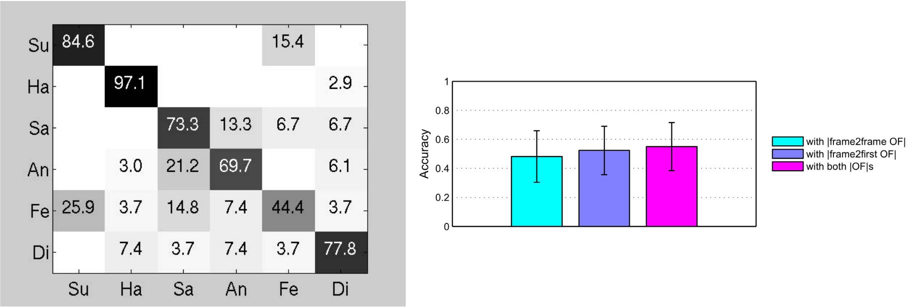


Fig. 7. *Left:* Average recognition accuracy on the Cohn-Kanade database plotted against the single image features and their combinations. The optical flow between the current and the first frame gives the best results, followed by the Gabor filter responses and the frame to frame optical flow. Best results are achieved by the combination of all three kinds of features. *Right:* Accuracy for each class label, as recognized from the tracks created thanks to the ground truth annotation (cyan bars on the left) and automatically extracted by the eye tracker (magenta, right).

Table 1 lists our results next the performance of other methods which used the Cohn-Kanade database and which published their recognition rates for each label. As can be seen, the results are comparable.

In an attempt to assess the contribution of each feature channel to the recognition, the left part of Figure 7 plots the accuracy achieved on the Cohn-Kanade database when each feature is used separately and in all their possible combinations. As can be seen,



	Su	Ha	Sa	An	Fe	Di
Su	84.6				15.4	
Ha		97.1				2.9
Sa			73.3	13.3	6.7	6.7
An		3.0	21.2	69.7		6.1
Fe	25.9	3.7	14.8	7.4	44.4	3.7
Di		7.4	3.7	7.4	3.7	77.8

Fig. 8. *Left:* confusion matrix for the MMI database. The higher rate of confusion with respect to the results obtained on the Cohn-Kanade database can be partly explained by the fact that manual annotations of the eye locations were not available. *Right:* Results obtained on the MMI database using the set of features originally employed by [12], with the addition of frame to frame and frame to first optical flow.

frame to first optical flow alone gives the best results, followed by log-Gabor responses and by optical flow computed between consecutive frames. The combination of all three features leads to the best results. In Figure 7, right, the performance for each class is plotted, depending on whether the tracks were extracted using the manual ground truth annotations of the eye locations (cyan bars on the left) or automatically, using the eye tracker (magenta, on the right). Results clearly worsen when the fully automatic method is employed, but not in the same extent for each class: surprise, happiness, and sadness are less affected by errors in the tracking than the other classes.

When training and testing on the MMI database, again in a 5-fold cross-validation fashion and with 200 patches of size $20 \times 50 \times 2$, we get the confusion matrix shown in Figure 8, left. There is a higher rate of misclassification compared to the results achieved on the Cohn-Kanade database, especially for fear. This could be partly explained by the fact that manual annotations of the eye centers were not available, but also by the lack of a precise annotation of the expression center in the sequences. Also, the expressions in the MMI database are more subtle than in the Cohn-Kanade dataset. On average, our method achieves a recognition rate of 76% on the MMI database and, as far as we know, we are the first ones to attempt at classifying the expressions directly (rather than Action Units) on this dataset. The right side of Figure 8 shows the average results obtained on the MMI sequences when using the features originally proposed by [12], with the addition of the two kinds of optical flow. The poor results of the original feature set serves as convincing support for the introduction of the log-Gabor filter responses, as explained in section 4.1

6 Conclusions

In this paper, we investigated the use of a Hough forest voting method for facial expression recognition. Our system extends previous work aimed at action recognition to the field of facial expression recognition, which are more subtle and hard to classify. We chose features encoding separately form and motion of the face, which allow us to capture the subtle differences in the facial expressions which a standard action recognition system could not. We evaluated the system on two standard databases, Cohn-Kanade and MMI, and achieved results comparable to the state of the art. Future work includes the investigation of additional features and the application of the method to the recognition of more naturalistic facial expression videos.

Acknowledgments. This work has been supported by funding from the Swiss National Foundation NCCR project IM2, the SNF project Vision-supported Speech-based Human Machine Interaction, as well as the EC project IURO. Angela Yao was also supported by funding from NSERC Canada.

References

1. Darwin, C.: The Expression of the Emotions in Man and Animals. John Murray (1872)
2. Ekman, P., Friesen, W.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology 17, 124–129 (1971)
3. Sebe, N., Sun, Y., Bakker, E., Lew, M., Cohen, I., Huang, T.: Towards authentic emotion recognition. In: International Conference on Systems, Man and Cybernetics (2004)
4. Cohn, J.F.: Foundations of human computing: facial expression and emotion. In: International Conference on Multimodal Interfaces, pp. 233–238 (2006)
5. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognition 36, 259–275 (2003)
6. Ekman, P., Friesen, W., Hager, J.: Facial action coding system: A technique for the measurement of facial movement (1978)
7. Ballard, D.H.: Generalizing the hough transform to detect arbitrary shapes. Pattern Recognition 13, 111–122 (1981)
8. Maji, S., Malik, J.: Object detection using a max-margin hough transform. In: Computer Vision and Pattern Recognition, pp. 1038–1045 (2009)
9. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: Computer Vision and Pattern Recognition (2009)
10. Ommer, B., Malik, J.: Multi-scale object detection by clustering lines. In: International Computer Vision Conference (2009)
11. Fanelli, G., Gall, J., Van Gool, L.: Hough transform-based mouth localization for audio-visual speech recognition. In: British Machine Vision Conference (2009)
12. Yao, A., Gall, J., Van Gool, L.: A hough transform-based voting framework for action recognition. In: Computer Vision and Pattern Recognition (2010)
13. Suwa, M., Sugie, N., Fujimora, K.: A preliminary note on pattern recognition of human emotional expression. In: International Joint Conference on Pattern Recognition (1978)
14. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. Trans. Patt. Anal. Mach. Intell. 31, 39–58 (2009)
15. Buenaposada, J.M., Muñoz, E., Baumela, L.: Recognising facial expressions in video sequences. Pattern Anal. Appl. 11, 101–116 (2008)
16. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: Machine learning and application to spontaneous behavior. In: Computer Vision and Pattern Recognition, pp. 568–573 (2005)
17. Aleksic, P.S., Katsaggelos, A.K.: Automatic facial expression recognition using facial animation parameters and multi-stream hmms. Trans. on Information Forensics and Security (1) (2006)

18. Dornaika, F., Davoine, F.: Simultaneous facial action tracking and expression recognition in the presence of head motion. Int. J. Comput. Vision 76, 257–281 (2008)
19. Shang, L., Chan, K.P.: Nonparametric discriminant hmm and application to facial expression recognition. In: Computer Vision and Pattern Recognition, pp. 2090–2096 (2009)
20. Essa, I., Pentland, A.: Coding, analysis, interpretation, and recognition of facial expressions. Transactions on Pattern Analysis and Machine Intelligence 19, 757–763 (1997)
21. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video. Transactions on Multimedia 8, 500–508 (2006)
22. Wu, T., Bartlett, M., Movellan, J.: Facial expression recognition using gabor motion energy filters. In: CVPR Workshop on Human Communicative Behavior Analysis (2010)
23. Shan, C., Gong, S., McOwan, P.: Facial expression recognition based on Local Binary Patterns: A comprehensive study. Image and Vision Computing 27, 803–816 (2009)
24. Zhao, G., Pietikäinen, M.: Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. Pattern Recogn. Lett. 30, 1117–1127 (2009)
25. Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J., Movellan, J.: Dynamics of facial expression extracted automatically from video. Image and Vision Computing 24, 615–625 (2006); Face Processing in Video Sequences
26. Lin, Z., Jian, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: International Computer Vision Conference (2009)
27. Reddy, K.K., Liu, J., Shah, M.: Incremental action recognition using feature-tree. In: International Computer Vision Conference (2009)
28. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
29. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: British Machine Vision Conference, pp. 47–56 (2006)
30. Valenti, R., Gevers, T.: Accurate eye center location and tracking using isophote curvature. In: Computer Vision and Pattern Recognition, pp. 1–8 (2008)
31. Schindler, K., Van Gool, L.J.: Action snippets: How many frames does human action recognition require? In: Computer Vision and Pattern Recognition (2008)
32. Field, D., et al.: Relations between the statistics of natural images and the response properties of cortical cells. Journal of the Optical Society of America A 4, 2379–2394 (1987)
33. Fukushima, K.: Neocognitron: a self-organizing neural network model for mechanisms of pattern recognition unaffected by shift in position. Biol. Cybernetics 36, 193–202 (1980)
34. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Automatic Face and Gesture Recognition, pp. 46–53 (2000)
35. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: International Conference on Multimedia and Expo, p. 5 (2005)
36. Lipori, G.: Manual annotations of facial fiducial points on the cohn kanade database, LAIV laboratory, University of Milan (2010), `http://lipori.dsi.unimi.it/download/gt2.html`