Analysis of HOG Suitability for Facial Traits Description in FER Problems

Marco Del Coco[1](✉), Pierluigi Carcagnì[1], Giuseppe Palestra[2], Marco Leo[1], and Cosimo Distante[1]

[1] National Research Council - National Institute of Optics, Arnesano, LE, Italy
marco.delcoco@ino.it

[2] Department of Computer Science, University of Bari, Bari, Italy

Abstract. Automatic Facial Expression Recognition is a topic of high interest especially due to the growing diffusion of assistive computing applications, as Human Robot Interaction, where a robust awareness of the people emotion is a key point. This paper proposes a novel automatic pipeline for facial expression recognition based on the analysis of the gradients distribution, on a single image, in order to characterize the face deformation in different expressions. Firstly, an accurate investigation of optimal HOG parameters has been done. Successively, a wide experimental session has been performed demonstrating the higher detection rate with respect to other State-of-the-Art methods. Moreover, an online testing session has been added in order to prove the robustness of our approach in real environments.

Keywords: Facial expression recognition · HOG · SVM

1 Introduction

Facial expression is one of the most common non-verbal way that humans use to convey internal emotion states and consequentially it plays a fundamental role in interpersonal interaction. Although there exists a wide range of possible face expressions, psychologists have identified six basic ones (happiness, sadness, fear, disgust, surprise, and anger) that are universally recognized [8]. It is straightforward that a system capable to perform an automatic recognition of the human emotion is a desirable task in the Human Computer Interaction (HCI) field (humanoid robots or digital signage applications). Unfortunately, the design of a system with an high recognition rate is a non trivial challenge, due to the subjects variability in terms of appearance and running expression.

A robust FER system should deal with the intrinsic variation of the same expression among different subjects in order to keep good performance with the unseen ones. Computer vision interest in the FER field has exponentially grown-up in the last years leading to a wide range of possible solutions.

There are two main approaches to FER; the first one uses image sequences while the second one is based on the analysis of a single image.

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 460–471, 2015.
DOI: 10.1007/978-3-319-23234-8_43

The use of image sequences means that many information are available for the analysis. Usually the neutral expression is used as the baseline face and then tracked in order to analyse the evolving expression over time [4,9,17]. Anyway, this approach shares a common lack: the dependence on a video sequence that evolves from the neutral expression to the expressive one. This constrain limits the use in real world environments where the evolution of facial expression is completely unpredictable. For this reason a more suitable solution for practical applications it to perform facial expression recognition on a single image. The approaches in literature that work on a single image can be conveniently categorized, depending on the strategies they use to lead to the recognition of the emotions, in two categories: Component Based Approaches and Global Approaches.

Component Based approaches preliminary extract some facial components and then try to classify the emotions on the basis of the matching among corresponding components or comparing the geometrical configuration among different components [7,14,19]. Unfortunately, even if in this kind of solutions the whole classification performances are not completely satisfactory due to the challenging alignment of components in different facial images (especially in case of extreme expressions) and to the computational burden if low-power systems are involved.

The above mentioned problems can be overcome by using "Global Approaches", i.e. approaches that directly try to extract a representation of the emotions from the appearance of the global face. This research area has been deeply investigated but there is still much effort to do, since it is very challenging to find a global set of descriptors able to robustly characterize human emotion traits. Many works exploiting most recent and reliable local descriptors have been proposed in recent years. Locally Binary Pattern (LBP) is used in [20] with kernel-based manifold approach, whereas directional information by the use of compass masks is exploited in [15]. Curvelet transform and online sequential extreme learning machine (OSELM) is instead proposed in [16].

As revealed in the previous discussion, FER problem is clearly related to the face deformation. Different persons could express the same emotion with some differences but the majority of the involved muscles work in such a way to give a coherent characterization of that emotions among different people of different ethnicity and gender.

This paper proposes a novel FER automatic pipeline based on the exploitation of the Histogram of Oriented Gradients (HOG) descriptor. It is a powerful shape description technique that counts occurrences of gradient orientations, in localized portions of an image, and that is intuitively useful to model the facial muscles shape by means of an edge analysis. To the best of our knowledge, HOG descriptors have been used as a tool for FER purposes only in [6] where authors just investigated the alignment perturbations for different descriptors and demonstrated that the best FER performances were obtained using LBP.

This paper proposes instead an in-depth analysis of how the HOG descriptors could be effectively exploited for facial expression recognition purposes and it demonstrates, by extensive experimental proofs, that the achieved FER performances outperform those of the leading state of the art approaches. Another important contribution of the paper is the introduction of an innovative algorithmic pipeline that takes as input a single facial image, performs a preliminary face detection and registration [1], apply the HOG descriptors and finally classify the facial expression by a group of Support Vector Machines (SVMs).

The rest of the paper is organized as in the followings: in section 2, the proposed methodology is detailed; in section 3 the optimization of the HOG parameters is experimentally performed; results, comparison against the state of the art approaches and tests on video streams are demanded to section 4. Conclusions are summarized in section 5.

2 Proposed Methodology

Facial expression recognition from generic images requires an algorithmic pipeline that involves different operative blocks. The scheme in Figure 1 has been used in this work: the first step detects the human faces in the image under investigation and then detected faces are cropped and registered [1]. These preliminary operations allow to get the quite similar position for eyes and in this way the subsequent HOG descriptor may be applied using a coherent spatial reference. The vector of features extracted by HOG is finally used for the classification of the facial emotions by SVM strategies. Each operative step is detailed in the following.

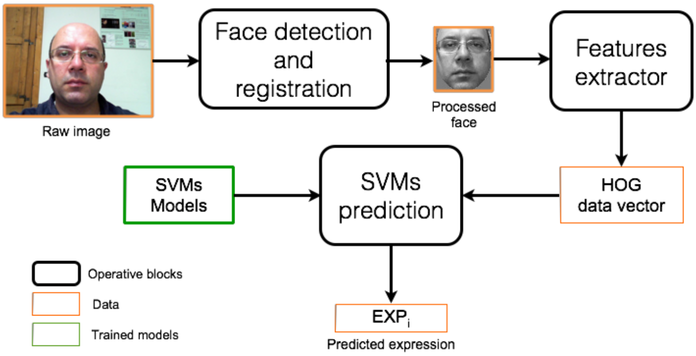


Fig. 1. Proposed system pipeline: faces are cropped and registered and then HOG descriptor is applied to build a data vector that is finally provided as input to a SVM bank that gives the estimation of the observed facial expression.

In *Face detection and registration* step human faces are detected in the input images and then registration and cropping operations are performed (Figure 2). Face detection makes use of both implicit and explicit knowledge: the explicit knowledge is based on the face geometry, color and appearance. On the other side, the implicit knowledge is integrated using the general object detection framework proposed by [18], which combines increasingly more complex classifiers in a cascade. Whenever a face is detected by a Viola Jones detector [18], it is fitted with an elliptical shape in order to rotate it to a vertical position. Successively a Viola-Jones based eye detector searches the eyes and exploits their position to scale the frontal face to a standard size of 65 × 59 pixels (registration).

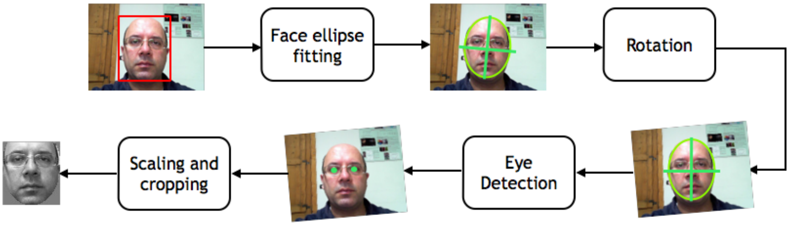


Fig. 2. Face registration: the detected face is fitted in an ellipse used to rotate the face in a perfectly vertical position; successively eyes are detected and used to scale the image and crop the zone of interest.

HOG descriptor is then applyed to the detected face region. Local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradients or edge positions. This statement led to the definition of the HOG technique that has been used in its mature form in Scale Invariant Features Transformation [11] and it has been widely exploited in human detection [3]. HOG descriptor is based on the accumulation of gradient directions over the pixels of a small spatial region referred as "cell" and in the subsequent construction of a 1D histogram whose concatenation supplies the feature vector to be considered for further purposes. Let L be the image to be analysed. The image is divided into cells of size $N \times N$ pixels (as in Figure 3 (a)) and the orientation $\theta_{x,y}$ of the gradient in each pixel is computed (Figure 3 (b-c)).Successively the orientations θ_i^j $i = 1...N^2$, i.e. belonging to the same cell j are quantized and accumulated in a M-bins histogram (Figure 3 (d-e)). Finally, all the achieved histograms are ordered and concatenated in a unique HOG histogram (Figure 3 (f)) that is the final outcome of this algorithmic step, i.e. the feature vector to be considered for the subsequent processing.

The feature vectors extracted by HOG descriptors are then given as input to a group of *Support Vector Machines* (SVMs). SVM is a discriminative classifier defined by a separating hyperplane. Given a set of labelled training data (supervised learning), the algorithm computes an optimal hyperplane (the trained model) which categorizes new examples in the right class. Anyway, such an approach is suitable only for a two classes problem whereas FER is a multiclass problem. It can be treated through the "one-against-one" [10]. Let k be the number of classes, then $k(k-1)/2$ classifiers are constructed where each one trains data from two classes. The final prediction is returned by a voting system

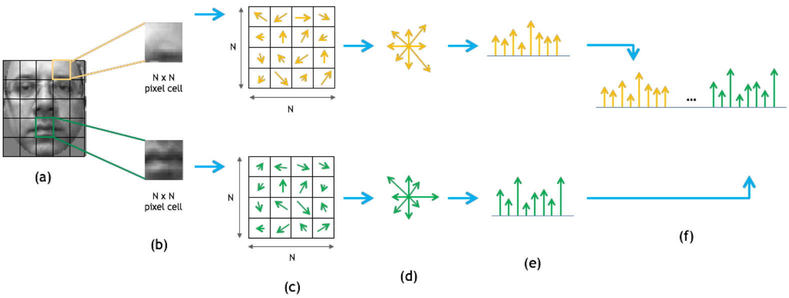


Fig. 3. HOG extraction features process: image is divided in cells of size $N \times N$ pixels. The orientation of all pixels is computed accumulated in an M-bins histogram of orientations. Finally, all cell histograms are concatenated in order to construct the final features vector. The example reports a cell size of 4 pixels and 8 orientation bins for the cell histograms.

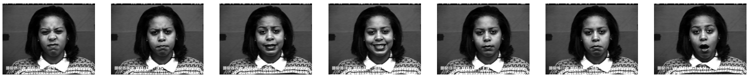
among all the classifiers. In particular, the multi C-support vector classification (multi C-SVC) learning task implemented in the LIBSVM library[2] was used in the following experiments. Radial Basis Function (RBF) was used as kernel with penalty parameter $C = 1000$ and $\gamma = 0.05$.

3 Experimental Setup

The evaluation of the proposed method has been performed on the Chon-Kanade dataset (CK+) [12], one of the most used to test the accuracy of FER solutions. It is made up by image sequences of people performing 6 facial expressions. Each sequence starts with a neutral face expression and ends with the expressive face.

In order to extract a balanced subset (quite the same number of instances for each considered expression) of images containing expressive faces, from the available sequences the following images were selected: the last image for the sequences related to the expression of anger, disgust and happiness; the last image for the first 68 sequences related to expression of surprise; the last and the fourth to the last images for the sequences related to the expression of fear and sadness. At the end, a set of 347 images was obtained with the following distribution among the considered classes of expressions: anger (45), disgust (59), fear (50), happiness (69), sadness (56) and surprise (68). An additional configuration of the previous subset was also introduced in order to test the accuracy performance with 7 classes and in this case 60 neutral faces were add to the aforementioned one. In Figures 4 some examples in the considered subsets of images are reported.

Once the datasets have been built up, the next step it to select the optimal value for the internal HOG parameters, i.e. the best configuration to capture the most discriminative information for the FER problem. HOG descriptor is characterized by two main parameters, the cell size and the number of orientation bins. Cell size represents the dimension of the patch involved in the single



(a) An (b) Di (c) Fe (d) Ha (e) Ne (f) Sa (g) Su

Fig. 4. Examples of expressions for the CK+ dataset. An=Anger, Di=Disgusted, Fe=Fearful, Ha=Happy, Ne=Neutral, Sa=Sad, Su=Surprised.

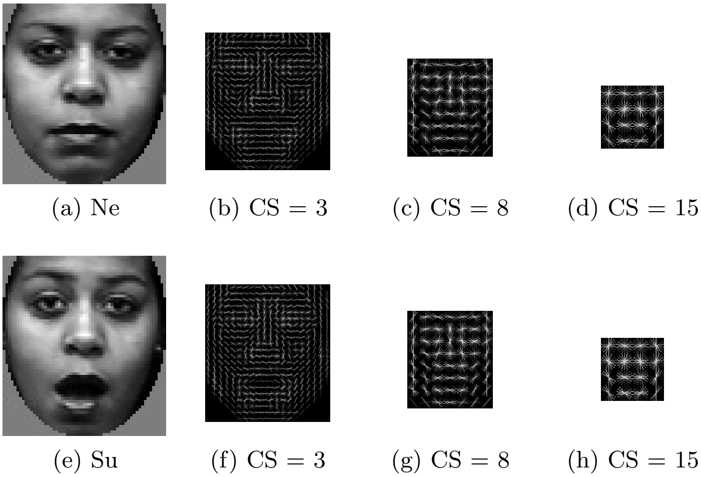
histogram computation. Using a large cell size the appearance information of a significant region is squeezed into a single cell histogram and then some details, useful for subsequent classification, can be lost. On the other hand, with a small cell size, high resolution analysis can be carried out, but this way the discrimination between useful and useless information could affect the classification step. The number of orientation bins refers instead to the quantization levels of the gradient information. A low number of orientations could drive to some loss of information and a consequent reduction in FER accuracy. On the contrary, an high number of quantization levels could spread-out the information along the bins, decreasing the FER accuracy as well. For these reasons, the choice of these parameters have to be carefully made taking into consideration the goal to be reached in the particular application context. How this choice was made for FER purposes is described in the following.

First of all, concerning cell size, a qualitative assessment can be made: in Figure 5 the normalized version of a neutral and a surprised face expressions is shown with the related processing outcomes obtained by HOG descriptor with a fixed number of 8 orientations and different values of cell size (3, 8 and 15 pixels). From figure could be deduced that the most discriminative representation is given, instead, by the use of a middle cell size (in the examples 8 pixels) whereas other cell size led to crowded bins distribution (3 pixels) or to a loss of correspondences between facial traits and HOG histogram (15 pixels).

The above qualitative evaluation can be also strengthened by a quantitative analysis of the FER accuracy sensitivity to both cell size and number of orientation bins. To perform this evaluation the proposed algorithmic pipeline was tested using a 10-fold cross validation with 12 possible values of the cell size (from 4 to 15 pixels) and different number of orientation bins (3, 5, 7, 9, 12, 15 and 55).

FER results for different numbers of orientation bins are graphically reported onto the y-axis in Figure 6 where the x-axis reports instead the cell size. From the figure it is possible to infer that a cell size of 7 pixels led to the best FER accuracy. Concerning the choice of the number of orientations, the best results were obtained with value set to 7 even if also with 9 or 12 orientations the FER accuracy did not change significatively.

Choosing the optimal parameter configuration (cell size of 7 pixels and 7 orientation bins) the proposed pipeline is able to correctly classify an average of 95.8% of the images supplied as input during the 10-fold cross validation process.



(a) Ne (b) CS = 3 (c) CS = 8 (d) CS = 15

(e) Su (f) CS = 3 (g) CS = 8 (h) CS = 15

Fig. 5. Examples of HOG (9 orientation) processing on normalised face images (Ne=Neutral, Su=Surprised.). CS is the cell size of the processed images.

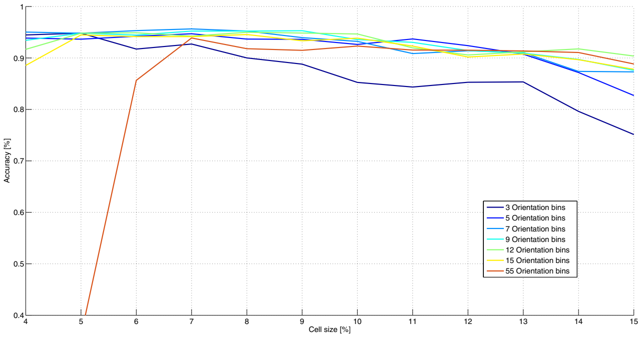


Fig. 6. FER results using different cell sizes and number of orientation bins for HOG descriptor (6 expressions): the x-axis report the cell size in pixel and the y-axis refers to the average accuracy percentage.

In order to verify that the best configuration of the selected HOG parameters keeps still valid also with different testing sets, the optimization carried out for the 6-expressions CK+ dataset has been extended to the the CK+ with 7 expressions. Results (showed in Figure 7) demonstrate that a cell size of 7 pixels and 7 orientation bins are the best configuration also for the 7-expressions CK+ dataset leading to a FER accuracy of 95.4%.

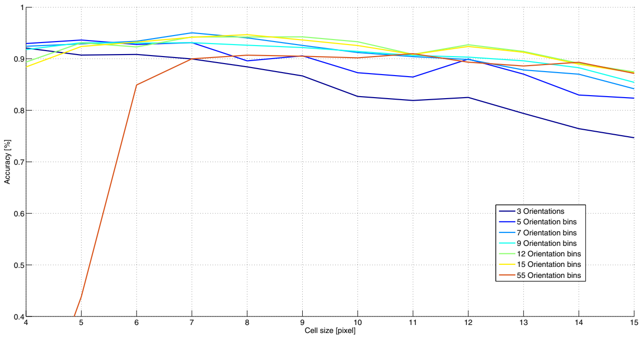


Fig. 7. FER results using different cell sizes and number of orientation bins for HOG descriptor (7 expressions): the x-axis report the cell size in pixel and the y-axis refers to the average accuracy percentage.

4 Experimental Results

In this section a wide experimental session is presented. Firstly, a detailed discussion of FER accuracy (in from of confusion matrices) is reported in subsection 4.1 whereas subsection 4.2 is aimed to the comparison with different techniques in the state of the art. Lastly, in subsection 4.3, an analysis of the behavior of the proposed pipeline when applied to video streams is presented.

4.1 Confusion Matrices for All the Datasets

Once established, in previous subsection, that there is a unique best configuration of the HOG parameters, the performance of the proposed approach are better analyzed by means of the confusion tables reported in Tables 2a and 2b. In particular, in a multi-class recognition problem as the FER one, the use of confusion tables makes possible a more detailed analysis of the results that can point out the missclassification cases and the interpretation of their possible causes. First of all, from tables it is possible to observe that, in the case of the CK+ dataset with 6 expressions, the accuracy was of 95.8% but, as expected, after the addition of the neutral expression it decreased to 95.4% (table 2b) due to the consequently increase of problem complexity.

These are very encouraging results considering the challenging benchmark used for testing.

Going into details, Tables 2a and 2b highlight the ambiguity between anger, disgusted and sad expressions. This becomes quite reasonable if the examples in Figure 4 are observed: for all aforementioned expressions, strict lips and low position of eyebrows are in fact very similar, in both location and appearance. For the same reasons, the neutral expressions introduced additional missclassifications in table 2b. Similarly, the sad expression experimented some erroneous

Table 1. Accuracy performance for proposed approach. (orientation bins = 7, cell size = 7). Ne=neutral, An=Anger, Di=Disgusted, Fe=Fearful, Ha=Happy, Sa=Sad, Su=Surprised.

	An	Di	Fe	Ha	Sa	Su
An	**88.6**	4.5	2.4	0	4.5	0
Di	5.5	**89.1**	1.8	1.8	0	1.8
Fe	0	0	**100**	0	0	0
Ha	0	0	0	**100**	0	0
Sa	0	0	0	0	**100**	0
Su	1.3	0	1.3	0	0	**97.4**

(a) CK+ (6-expressions): average accuracy = 95.8%

	Ne	An	Di	Fe	Ha	Sa	Su
Ne	**89.6**	1.8	0	0	0	8.6	0
An	4.4	**86.8**	4.4	0	0	4.4	0
Di	0	5.4	**92.9**	1.7	0	0	0
Fe	0	0	0	**93.9**	4.1	0	2.0
Ha	0	0	0	0	**100**	0	0
Sa	0	0	1.8	0	0	**98.2**	0
Su	1.3	0	0	1.3	0	0	**97.4**

(b) CK+ (7-expressions): average accuracy = 95.4%

classification in the anger face expression due to the strict lips and low position of eyebrows that are very similar for the two expressions.

4.2 Comparison with the State of the Art

In this subsection the proposed pipeline is compared with the leading State-of-the-Art solutions in literature. In order to proceed in the fairest way, the comparison was performed with all those solutions that used an evaluation protocol based on the CK+ dataset with 6 expressions. Table 2 reports the comparison results demonstrating that the proposed approach gave the best average recognition rate. In particular it is worth noting that the performance achieved by the proposed approach exceed also those of the recent work in [7] that represents the reference point for the FER problem. A deeper analysis of the Table 2 evidences that the proposed method suffers more than competitors to recognize the expression of disgust. This drawback could be due to the fact that, while performing this expression, the facial muscles shape is quite similar to that of the expression of anger (see Figure 4) then the edge analysis performed by HOG

Table 2. Performance comparison of our approach versus different State-of-the-Art approaches (CK+ 6 expressions). An=Anger, Di=Disgusted, Fe=Fearful, Ha=Happy, Sa=Sad, Su=Surprised.

	[17]	[14]	[19]	[7]	PROPOSED
An	82.5	87.1	87.1	87.8	88.6
Di	97.5	91.6	90.2	93.3	89.1
Fe	95.0	91.0	92.0	94.3	100
Ha	100	96.9	98.1	94.2	100
Sa	92.5	84.6	91.5	96.4	100
Su	92.5	91.2	100	98.5	97.4
AV	93.3	90.4	93.1	94.1	**95.8**

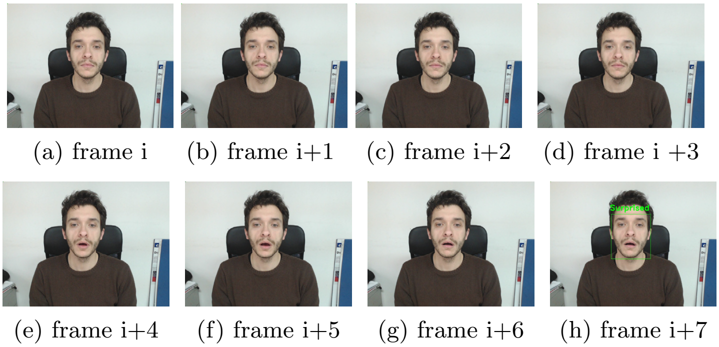
sometimes cannot be able to bring to light differences as other approaches based on texture analysis can instead highlight. However, this is a limitation only for the recognition of the expression of disgust since for all the remaining expressions the FER performances of the proposed method largely exceed those of the comparing methods highlighting that the analysis of the edges is the best method for the FER problem.

4.3 Tests on Video Streams

The experiments reported in previous subsections were relative to the recognition of facial expressions in a single image containing a clearly defined expression. In common application contexts, the automatic systems have to perform FER by analyzing image sequences in which not all the image contain a clear expression, or where there are transitions between expressions. This section aims thus at analyzing the behavior of the proposed pipeline when applied to a continuous video streaming.

To make the system suitable for video streams analysis, a decision making strategy based on the temporal consistency of FER outcomes has been introduced. The decision about the expression in a video is taken by analyzing a temporal window of size m and verifying if at least n ($n < m$) frames in the window are classified as containing the same facial expression. In the experiments the following setting was used: $n = 4$, $m = 5$,

Subjects of different gender and age have been involved in the experiment. More specifically, all subjects were asked to sit in front of the camera and perform, without a particular order, some of the aforementioned expressions. It is worth noting that, in this way, no constraints, about the passage from the neutral expression to the expressive one have been introduced, leaving the tester free to perform every possible transition among different expressions. The testing system was configured with a webcam with a resolution of 640 × 480 pixels



(a) frame i (b) frame i+1 (c) frame i+2 (d) frame i +3

(e) frame i+4 (f) frame i+5 (g) frame i+6 (h) frame i+7

Fig. 8. Example of expression detection performed by the proposed system: the expression evolves over time; once the decision making rule is satisfied, the prediction is printed out.

and a PC customized by an i5 processor (2,66 GHz) and 4 GB RAM where the system worked in real-time.

The system, evaluated from a qualitative point of view, exhibited a good capacity to recognize all the emotions performed by the testing subjects with a quite low presence of false positive thanks to the filtering performed by the temporal windows and the decision rule above mentioned. One example of the system output is reported in Figure 8. The subject performs a particlular expression that is recognized and showed when a sufficient number of similar classifications are counted in the time window.

5 Conclusions

In this paper, a novel FER automatic pipeline, based on the exploitation of the Histogram of Oriented Gradients (HOG) descriptor, has been proposed. An in-depth analysis, of how the HOG descriptor can be effectively exploited for facial expression recognition purposes, has been supplied. Extensive experiments on CK+ publicly available dataset have been carried out demonstrating that the achieved FER performances outperform those of the leading state of the art approaches. Finally, additional experiments on video streams demonstrated the suitability of the proposed approach for real application contexts. Future works will deal with the test of the presented system in the field of assistive technologies. A social robot will be equipped with the proposed FER solution in order to acquire the awareness about the emotional state of an interacting subject and consequentially adopt an adequate reaction. An additional step will be provided by the coupling with face recognition and re-identification strategies [5,13] aided to keep a temporal consistence of the emotion of different subject among successive interaction sessions.

References

1. Castrillón, M., Déniz, O., Guerra, C., Hernández, M.: Encara2: Real-time detection of multiple faces at different resolutions in video streams. Journal of Visual Communication and Image Representation **18**(2), 130–140 (2007)
2. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(3), 27 (2011)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893 (2005)
4. Dornaika, F., Lazkano, E., Sierra, B.: Improving dynamic facial expression recognition with feature subset selection. Pattern Recognition Letters **32**(5), 740–748 (2011)
5. Farinella, G.M., Farioli, G., Battiato, S., Leonardi, S., Gallo, G.: Face reidentification for digital signage applications. In: Distante, C., Battiato, S., Cavallaro, A. (eds.) VAAM 2014. LNCS, vol. 8811, pp. 40–52. Springer, Heidelberg (2014)
6. Gritti, T., Shan, C., Jeanne, V., Braspenning, R.: Local features based facial expression recognition with face registration errors. In: 8th IEEE International Conference on Automatic Face Gesture Recognition, FG 2008, pp. 1–8 (2008)

7. Happy, S., Routray, A.: Automatic facial expression recognition using features of salient facial patches. IEEE Transactions on Affective Computing **PP**(99), 1–1 (2015)
8. Izard, C.: The face of emotion. Century psychology series. Appleton-Century-Crofts (1971)
9. Khan, R.A., Meyer, A., Konik, H., Bouakaz, S.: Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recognition Letters **34**(10), 1159–1168 (2013)
10. Knerr, S., Personnaz, L., Dreyfus, G.: Single-layer learning revisited: a stepwise procedure for building and training a neural network. Neurocomputing **68**, 41–50 (1990)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
12. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
13. Martiriggiano, T., Leo, M., D'Orazio, T., Distante, A.: Face recognition by kernel independent component analysis. In: Ali, M., Esposito, F. (eds.) IEA/AIE 2005. LNCS (LNAI), vol. 3533, pp. 55–58. Springer, Heidelberg (2005)
14. Poursaberi, A., Noubari, H., Gavrilova, M., Yanushkevich, S.: Gauss–laguerrewavelet textural feature fusion with geometrical information for facial expression identification. EURASIP Journal on Image and Video Processing **2012**(1), 17 (2012)
15. Rivera, R., Castillo, R., Chae, O.: Local directional number pattern for face analysis: Face and expression recognition. IEEE Transactions on Image Processing **22**(5), 1740–1752 (2013)
16. Uddin, M., Lee, J., Kim, T.S.: An enhanced independent component-based human facial expression recognition from video. IEEE Transactions on Consumer Electronics **55**(4), 2216–2224 (2009)
17. Uçar, A., Demir, Y., Güzeliş, C.: A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering.Neural Computing and Applications, 1–12 (2014)
18. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision **57**(2), 137–154 (2004)
19. Zhang, L., Tjondronegoro, D.: Facial expression recognition using facial movement features. IEEE Transactions on Affective Computing **2**(4), 219–229 (2011)
20. Zhao, X., Zhang, S.: Facial expression recognition based on local binary patterns and kernel discriminant isomap. Sensors **11**(10), 9573–9588 (2011)