Hand-gesture recognition : comparative study of global, semi-local ans local approaches

Jean-François Collumeau, Rémy Leconge, Bruno Emile, Hélène Laurent.

Hand-gesture recognition: comparative study of global, semi-local and local approaches

Jean-François Collumeau*, Rémy Leconge†, Bruno Emile‡, Hélène Laurent*

*ENSI de Bourges - Laboratoire PRISME
88 boulevard Lahitolle, 18020 Bourges Cedex, France
jeanfrancois.collumeau@ensi-bourges.fr
helene.laurent@ensi-bourges.fr

†Polytech'Orléans - Laboratoire PRISME
12 rue de Blois - BP 6744, 45067 Orléans cedex 02, France
remy.leconge@univ-orleans.fr

‡IUT de l'Indre - Laboratoire PRISME
2 avenue F. Mitterrand, 36000 Châteauroux, France
bruno.emile@univ-orleans.fr

***Abstract*—Asepsis preservation in operating rooms is essential for limiting patient infection by hospital-acquired diseases. For health reasons, surgeons may not be directly in contact with sterile equipment surrounding them, and must instead rely on assistants to interact with these in their place. Gesture-based Human-Computer Interfaces constitute an interesting option for allowing surgeons to control such equipment without breaking asepsis-preservation rules.**

In this paper, we describe the results obtained by a preparatory study to such a system. Methods representative of the three main approaches in object description (i.e. local, semi-local and global approaches) are compared in terms of hand posture recognition performance when picture backgrounds are not removed. A picture database presenting variations in the postures' angles and illumination, as well as plain and altered backgrounds was created to this end. We study the effects of incomplete training databases, user-dependency and altered backgrounds on recognition rates.

Obtained results give semi-local approaches such as Dalal et al.'s Histograms of Oriented Gradients more promising than local or global ones. However, performance degradation when considering non-uniform picture backgrounds makes pre-treatments aiming at segmenting the hand from the background appear to be necessary.

I. Introduction

During surgical interventions, surgeons are not allowed to interact directly with the entire equipment of the operating room (OR) because of asepsis preservation measures. As illustrated by figure 1, this equipment is becoming more and more complex and varied, including mobile lights, various screens - fixed or mobile - presenting information from the patient file or images collected online by medical tools, fixed cameras... In order to place computer screens closer, change visualization modes, modify the operating table's tilt or decrease ambient lights..., surgeons usually appeal to a nurse who is in charge of interpreting the expressed instructions. This protocol is particularly awkward. An objective of the *CORTECS* project, which involves different industrial partners, focuses on improving the working comfort for the hospital team and particularly for surgeons. This includes among others a remote non-contact OR equipment control designed for surgeons and their assistants. In opposition to voice-controlled systems as proposed by [8], computer vision-based systems are insensitive to the OR's noisy environment. Because surgeons appear to be able to free one hand easily for non-urgent commands, the foreseen system is intended to be used in non-urgent situations during both pre-operative and operative stages. It will enable users to command equipments, choose the global positions and orientations of mobile devices and perform finer adjustments. Hand-based commands will be issued by performing various hand postures. The final system should allow surgeons to remain close to the patient while avoiding distractions, i.e. surgeon attention decrease due to formulating instructions. Most of the actual hand-based surgeon-computer interfaces deal with the development of remote control systems for browsing through patients MRI or CT scan images [1], [2], [3]. They require surgeons to stand close to the main control wall, or at least to be positioned in front of the controlled device, allowing relatively low disparities in acquisition points of view and lighting conditions. We plan to make our application tolerant to rotations of the hand postures proposed as inputs as well as to illumination changes, without perturbing the system's performances. Coupled with image acquisition from multiple cameras spread in the OR, this would free surgeons from the forementionned positionning constraint.

The study presented in this paper is a preliminary one focusing on hand posture recognition, which will be one of the various steps constituting the complete process. This paper's objective is to perform a comparative study of classical tools commonly used for object recognition and check if these techniques can be extended to the specific context considered here. The disparity between our objects (which are always complete hands, differing only slightly in finger positioning)

will indeed be much smaller in the considered situations than in usual object recognition, such as seen in content-based image retrieval for example. Three main approaches can be found in the literature for object recognition. The so-called global approaches rely on the computation of invariant descriptors on an entire image containing the object. Semi-local approaches split the object image into several blocks and compute descriptors on each block. Finally, local approaches use the same kind of descriptors in combination with keypoint detectors, calculating features in the neighborhood of each detected keypoint. In order to check which of these techniques would be the most adequate for our application, we selected one representative method for each approach among the most used in the literature before comparing their relative performances on a home-made database.

The next section presents the selected methods for each of the global, semi-local and local approaches. Section 3 is dedicated to the comparative study itself: we detail the created image database and comment the obtained results. The recognition being based on learning techniques, the influence of the learning database size on each approach's performances has been studied as well. Their robustness faced with different image background alterations is also evaluated. Finally, some conclusions and perspectives are drawn from these results.

Fig. 1. Operating theater presented in MAQUET showroom, Rastatt, Germany.

II. Tested approaches

A. Global Approach (Zernike)

In order to characterize an object that can appear at different orientations and scales in an image, an invariant descriptor must be used. Some invariant descriptors in the literature such as, Hu moments, Fourier Mellin transform and Zernike moments, have been compared in a previous work [4]. The conclusion is that Zernike moments gives better results. Zernike moments are computed from a set of Zernike polynomials. This set is complete and orthonormal in the interior of the unit circle. These 2D image moments allow to overcome the major drawbacks of regular geometrical moments regarding noise effects and presence of image quantization error. Their orthogonality property helps in achieving a near zero value of redundancy measure in a set of moments functions. The Zernike moments formulation is given below [5]:

$$A_{mn} = \frac{m+1}{\pi} \sum_x \sum_y I(x,y)[V_{mn}(x,y)] \qquad (1)$$

with $x^2 + y^2 < 1$.
The values of m and n define the moment order and $I(x,y)$ is the pixel intensity of the image I. Zernike polynomials $V_{mn}(x,y)$ are expressed in the radial-polar form:

$$V_{mn}(r,\theta) = R_{mn}(r)e^{-jn\theta} \qquad (2)$$

where $R_{mn}(r)$ is the radial polynomial given by:

$$R_{mn}(r) = \sum_{s=0}^{\frac{m-|n|}{2}} \frac{(-1)^s (m-s)! r^{m-2s}}{s!(\frac{m+|n|}{2}-s)!(\frac{m-|n|}{2}-s)!} \qquad (3)$$

These moments yield invariance with respect to translation, scale and rotation. Zenike moments have been used by [13] for recognizing hand postures with a multivariate decision tree-based classifier in a human-robot interaction context. In this study we combine these descriptors with a linear SVM.

B. Semi-local approach (HOG)

Histogram of Oriented Gradient descriptors provide a dense indeed overlapping description of image regions. The local shape information is captured by encoding image gradients orientations in histograms. The first step of calculation is the computation of the gradient values. The second step of calculation involves creating the cell histograms. Each pixel within the cell casts a weighted vote for an orientation-based histogram channel based on the values found in the gradient computation. In order to account for changes in illumination and contrast, the gradient strengths must be locally normalized, which requires grouping the cells together into larger, spatially-connected blocks. The last step concerns block normalization. Dalal et al. [7] have proposed Histogram of Oriented Gradients in the case of human detection combined with a linear SVM. Lu et al. [10] and Kaniche et al. [11] used temporal HOGs for action categorization and gesture recognition respectively.

C. Local approach (SIFT)

Sift is a well known descriptor and has been largely used and studied since Lowe created it in 1999. It allows to detect and extract features which are invariant to image scaling and rotation, and partially invariant to changes in illumination and 3D camera viewpoint. They are well localized in both the spatial and frequency domains, reducing the probability of disruption by occlusion, clutter, or noise. The SIFT descriptor, as described in [6], consists of four stages:

- scale-space peak selection: searching over all scales and image locations, it is efficiently implemented by using a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation,
- interest point localization: among all keypoint candidates, the ones that have low contrast and low stability are rejected,
- orientation assignment: one or more orientations are assigned to each keypoint based on local image gradient directions. Invariance to image rotation is achieved by

performing all future operations relative to this orientation.

- descriptor: the descriptor is computed from a 4*4 location Cartesian grid. The gradient on each location bin is computed on the patch around the keypoints and is then quantized into 8 orientation bins. This leads to a 128-element vector.

The three first stages correspond to the localization of keypoints. The SIFT descriptor gives good results in the case of object recognition when it can find relevant keypoints. Compared to global descriptors, it is quite robust to partially occluded objects. Wang et al. [12] combined SIFT and boosting methods to their advantage for a human-robot interface. We used Rob Hess' SIFT C implementation [15] in combination with a linear SVM classifier in this paper.

III. Comparative study

A. Database

The image database we created contains 468 pictures. It consists in 6 hand postures selected as being easily reproducible: open palm, closed fist, Y posture, U posture, OK posture and thumbs up. They are presented in figure 2. This gesture vocabulary can of course be extended or customized according to the user's affinity.

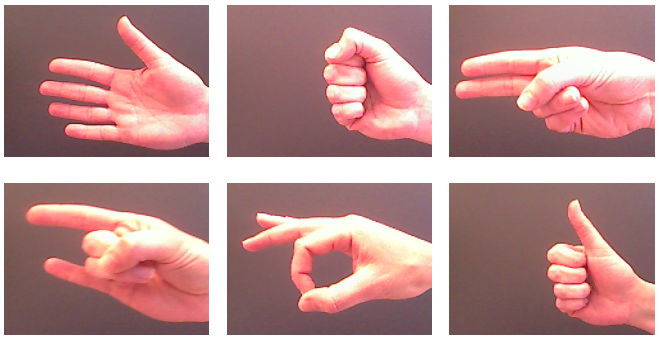
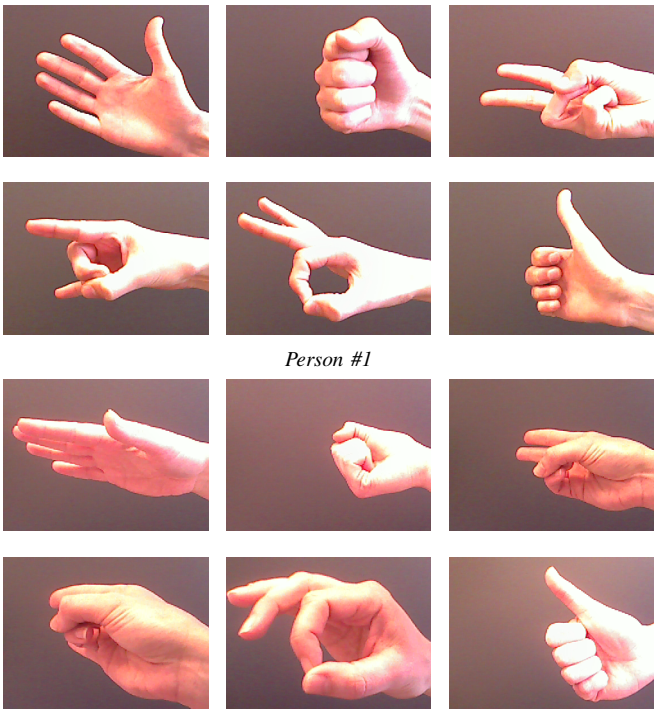


Fig. 2. Six postures constituting the gesture vocabulary

Two persons have participated in this experiment. For each posture, 26 views (320 $\times$ 240 pixels images) presenting orientation and illumination changes have been acquired. All of those present clean black backgrounds. As can be seen in figure 3, images corresponding to the same posture can be relatively disparate for the same "speaker", depending on the considered angle and on lighting conditions. They may also differ between "speakers": thumbs up for example corresponds to tucked fingers and an extended thumb for the first person while the second person closes the fist when extending the thumb; in the same manner, the U posture is made with spread-out fingers for the first person, while those are placed in a more parallel way by the second person.

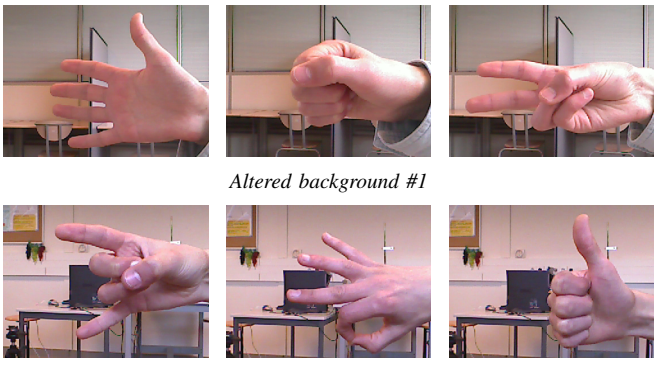
In order to check the robustness of the selected approaches, we acquired, for the first person and for each of the six postures, 26 images with two different altered backgrounds. These alterations insert shapes different than the vocabulary postures in the pictures' background (see figure 4).



Person #1

Person #2

Fig. 3. Postures examples with different orientations and illuminations



Altered background #1

Altered background #2

Fig. 4. Postures examples with altered backgrounds

B. Experiments

As our work aims at eventually enabling each surgeon to specify his command postures, specific training for each surgeon will be necessary. Having surgeons perform a small amount of training postures would be preferable in order for the training step not to be long and cumbersome for them. We have therefore studied the possibility of training a Support Vector Machine (SVM) classifier on only part of each posture's set of pictures. Due to the low ratio between the database's size and the amount of descriptors involved for describing each of the pictures (128 descriptors for Zernike moments, about 5000 and 4000 for HOG and SIFT respectively) we selected a linear kernel for the SVM according to Chang and Lin's [14] recommandations on kernel selection. Two

distinct experiments were led in order to check respectively the influence of a partial training set and of background alterations on recognition rates.

First, we proceeded in the following manner for determining the influence of an incomplete training : we began by selecting randomly a fixed number of training samples out of the first speaker's database and training the SVM classifier with them. Remaining pictures (that is, pictures from the speaker's database which were not selected for training) were then assigned a posture label by the SVM. The considered number of samples varied from 3 (very partial training) to 21 (almost complete training) for each posture. This experiment was carried three times, once with each speaker separately and once on a 52 images per posture database consisting of both users' pictures mixed in a single database.

In a second part, the background influence was studied by training the SVM classifier on a single-speaker database presenting a clean black background. Pictures presenting postures performed by the same speaker on each of the two altered backgrounds were then classified by the SVM.

C. Results and discussion

Partial training experiment outcomes are shown in tables I, II and III; results for background influence are presented in table IV.

Results obtained from sampling images from the database in order to perform a partial training unsurprisingly show an increase in recognition performance when the amount of training samples increases. However, maximum performance is never achieved when training isn't performed on a complete database. Recognition rates for both speakers are respectively shown in tables I and II. Speaker 2's recognition rates are slightly lower than speaker 1's because it appears rotation angles were lower in the first speaker's postures variations, hence leading to less differentiated postures and easier classification.

Training images	*Training sample ratio*	*Zernike*	*HOG*	*SIFT*
3	12%	39,40%	60,71%	33,04%
6	23%	51,00%	71,17%	50,36%
9	35%	56,75%	75,17%	54,13%
12	46%	61,79%	80,48%	61,18%
15	58%	66,07%	84,23%	66,04%
18	69%	70,21%	86,04%	68,34%
21	81%	73,32%	87,66%	69,32%

TABLE I
PARTIAL TRAINING INFLUENCE - SPEAKER #1

Performance ranking gives HOG as the best-performing method, followed by Zernike moments. However the gap between recognition rates achieved by SIFT and Zernike moments appears to be highly dependant on the speaker performing the postures, as can be seen in figure 5. As small as this gap may be, SIFT nevertheless ranks last, which is surprising given its usual prevalence over other descriptors in the literature. We believe SIFT's poor performances come from the fact that hand postures are not as dissimilar as distinct objects, which were

Training images	*Training sample ratio*	*Zernike*	*HOG*	*SIFT*
3	12%	39,44%	47,40%	29,86%
6	23%	51,73%	60,00%	40,50%
9	35%	56,77%	65,19%	43,34%
12	46%	62,27%	67,38%	51,68%
15	58%	64,04%	72,23%	51,52%
18	69%	68,55%	70,85%	52,52%
21	81%	69,33%	75,33%	58,74%

TABLE II
PARTIAL TRAINING INFLUENCE - SPEAKER #2

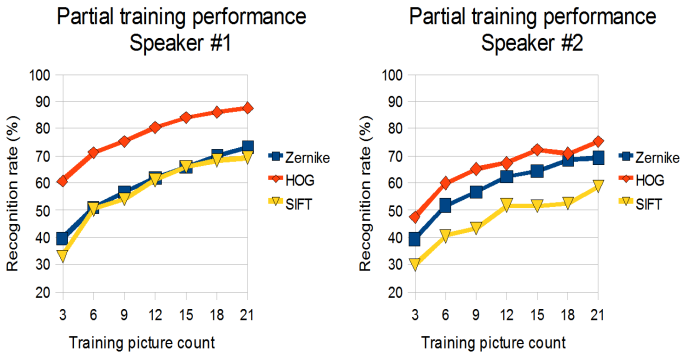


Fig. 5. Evolution curves of performance versus training database size

SIFT's initial application. In fact, some postures differ only slightly; for example, an open palm posture and a thumbs up differ only by half of the hand. Pictures of both postures, as well as the SIFT keypoints calculated on them, can be seen in figure 6's top row. The bottom row shows another example of keypoints commom to the closed fist and U postures. The size and orientation of the arrows present on these pictures show the keypoints importance and orientation. The keypoints' positions are located at the origin of each arrow. One can easily see most keypoints of lesser importance and some of importance differ from one posture to another, but a non-negligible amount of significant keypoints are also shared by these postures. This keypoint-sharing between postures, which is less influent in standard object retrieval, is likely to account for SIFT's poor posture recognition rate.

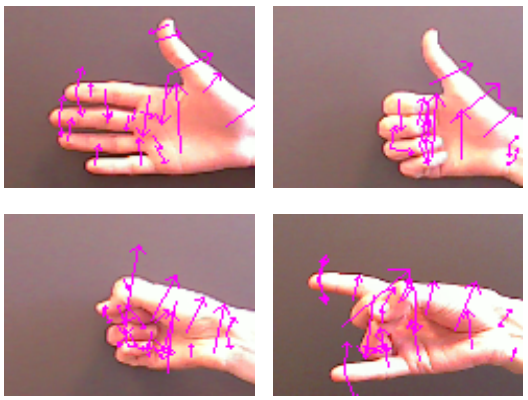


Fig. 6. Examples of postures presenting shared SIFT keypoints

Training the SVM on both speakers shows a decrease in recognition rates compared to single-speaker recognition, as can be seen in table III. This is partly due to a decrease in the ratio between training samples and the picture database's size. Indeed, this means training is never almost complete, but at best done with 40% pictures as training samples, versus 81% in the single speaker case. Nevertheless, comparing slightly equal training ratios emphasizes the fact Zernike and HOG descriptors suffer from a loss of performance when training on more than one unique speaker. The SIFT method is mainly unaffected by a multiple-speaker partial training.

Training images	*Training sample ratio*	*Zernike*	*HOG*	*SIFT*
3	6%	30,88%	45,23%	27,70%
6	12%	35,73%	57,50%	41,98%
9	17%	38,10%	61,28%	47,22%
12	23%	39,22%	63,51%	51,78%
15	29%	41,88%	65,74%	53,80%
18	35%	43,97%	68,47%	54,72%
21	40%	45,43%	70,77%	56,52%

TABLE III
PARTIAL TRAINING INFLUENCE - MIXED SPEAKERS

The second part of our experiments was focused on checking the influence of a non-uniform background on the recognition rates. Obtained results are summarized in table IV. It is obvious recognition rates are highly affected by altering the pictures' background, dramatically decreasing all descriptors' performances. As previously, the HOG descriptor performs best when facing postures on altered background, widely outweighting the other methods by a scale of two to three. The SIFT method gives second-best results, followed by Zernike moments.

Training images	*Zernike*	*HOG*	*SIFT*
Altered BG #1	10,25%	44,87%	26,92%
Altered BG #2	12,82%	33,33%	19,23%

TABLE IV
BACKGROUND INFLUENCE ON RECOGNITION RATES

Our experiments show the HOG method as yielding the best overal performances, outweighting both SIFT and Zernike moments in recognition rates. Although Zernike moments slightly outperform SIFT when using an incomplete training database, they remain far more sensitive to non-uniform backgrounds and user-dependant posture variations than SIFT.

IV. CONLUSION

In the end, HOG stands out as the best method for recognizing hand postures given a complete picture as input under both partial training and altered backgrounds constraints. Zernike moments proved to be competitive only in absence of altered backgrounds. SIFT descriptors performed below-average in hand posture recognition, which might be due to some significative keypoints being shared by different postures, but remained unaffected by multiple-speaker recognition unlike both previous methods. Hence, according to representative methods of each approach, semi-local approaches such as HOG appear the most promising for describing hand postures independently from the picture's background. However, achieved recognition rates remain far too low for the development of a reliable gesture-based human-computer interface, especially in situations involving medical care.

In order to achieve higher recognition rates, future works will include extending our study to posture pictures where the background has been removed and to contours of such hand postures. In the latter case, Fourier-based descriptors will be examined as well. Feature selection through PCA and boosting techniques prior to training and recogntion steps might improve recognition rates, especially with the SIFT method by helping in the removal of features shared by different postures. Vote-based label decision on several time-successive pictures might also help achieving better performances. The picture database will be extended with pictures taken in real OR environments in order to take into account the true conceivable magnitude of lighting changes in the OR. We also plan debating with surgeons about their preferences in terms of postures to be used.

ACKNOWLEDGMENT

The authors would like to thank the Regional Council of the Centre and the French Industry Ministry for their financial support within the framework of the Cortecs project through the Competitiveness Pole S2E2.

REFERENCES

[1] J. P. Wachs, H. I. Stern, Y. Edan, M. Gillam, J. Handler, C. Feied and M. Smith, *A gesture-based tool for sterile browsing of radiology images*, Journal of the American Medical Informatics Association, Volume 15, Number 3, 2008.

[2] T. Kipshagen, M. Graw, V. Tronnier, M. Bonsanto and U. G. Hofmann, *Touch- and marker-free interaction with medical software*, World Congress on Medical Physics and Biomedical Engineering, Volume 25, Springer, 2009.

[3] D. L. M. Achacon, D. M. Carlos, M. K. Puyaoan, C. T. Clarin and P. C. Naval, *REALISM: Real-time hand gesture interface for surgeons and medical experts*, 2009.

[4] A. Choksuriwong, B. Emile, C. Rosenberger and H. Laurent, *Comparative study of global invariant descriptors for object recognition*, Journal of Electronic Imaging, Volume 17, Issue 2, 023015, 2008.

[5] A. Khotanzad and H. Hong, *Invariant image recognition by Zernike moments*, IEEE Transactions on Pattern Analysis and Machine Intelligent, Volume 12, Issue 5, pp. 489-497, 1990.

[6] D. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, Volume 60, Issue 2, pp. 91-110, 2004.

[7] N. Dalal, B. Triggs, C. Schmid, *Human detection using oriented histograms of flow and appearance*, European Conference on Computer Vision (ECCV), 2006.

[8] Riisgard, J.E. Bardram, *ActiveTheatre A Collaborative, Event-Based Capture and Access System for the Operating Theatre*, UbiComp 2005: Ubiquitous Computing (2005), pp. 375-392.

[9] J.S. Chang, E.Y. Kim, H.J. Kim, *Mobile robot control using hand-shape recognition*. Transactions of the Institute of Measurement and Control June 2008 30: 143-152.

[10] W.L. Lu, J.J. Little, *Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor*. Computer and Robot Vision (CRV) 2006.

[11] M.-B. Kaniche, F. Brmond, *Tracking hog descriptors for gesture recognition*. AVSS09: 6th IEEE International Conference on Ad- vanced Video and Signal Based Surveillance, 2009.

[12] C.C. Wang, C.C. Wang, *Hand Posture Recognition Using Adaboost with SIFT for Human Robot Interaction*. Proceedings of the International Conference on Advanced Robotics (ICAR), 2007.
[13] L. Gu, J. Su, *Natural hand posture recognition based on Zernike moments and hierarchical classifier*. IEEE International Conference on Robotics and Automation (ICRA), 2008.
[14] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*. 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
[15] Rob Hess, *An open-source SIFTLibrary*. Proceedings of the international conference on Multimedia, 2010.