Hand Detection in First Person Vision

Pietro Morerio, Lucio Marcenaro, Carlo S. Regazzoni
DITEN
University of Genoa
Via Opera Pia 11A 16145 Genoa - Italy
Email: {pmorerio, mlucio}@ginevra.dibe.unige.it

Abstract—The emergence of new pervasive wearable technologies (e.g. action cameras and smart glasses) calls attention to the so called First Person Vision (FPV). In the future, more and more everyday-life videos will be shot from a first-person point of view, overturning the classical fixed-camera understanding of Vision, specializing the existing knowledge of moving cameras and bringing new challenges in the field of video processing. The trend in research is going to be oriented towards a new type of computer vision, centred on moving sensors and driven by the need for new applications for wearable devices. We identify hand tracking and gesture recognition as an essential topic in this field, motivated by the simple realization that we often look at our hand, even while performing the simplest tasks in everyday life. In addition, the next frontier in user interfaces are hands-free devices.

In this work we argue that applications based on FPV may involve information fusion at various complexity and abstraction levels, ranging from pure image processing to inference over patterns. We address the lowest, by proposing a first investigation on hand detection from a first-person point of view sensor and some preliminary results obtained fusing colour and optic flow information.

I. INTRODUCTION

Portable head-mounted action cameras, built for recording dynamic high-quality first-person videos, have become a common item among many sportsmen, and have lately aroused the interest of researchers [1]. Such cameras allow to record first-person point-of-view experiences, nothing more.

Last year Google announced Project Glass and started publishing short teasers on the web. Its goal was to build a wearable computer (Figure 1) that records your perspective of the world, and also delivers information through a head-up display. The idea is not completely new, as back in 2004 Pentland et al. [2] proposed a wearable data collection system called InSense. The concept of “wearable” has surely evolved by then.


A developer edition of Google Glasses is ready by now, and the hashtag “#IfIHadGlass” was lately flooding Twitter and Google+. Unfortunately only American citizens can apply for getting a prototype, and the final version is not due until 2014 at the earliest. Microsoft obviously set itself off in the chase and got a patent on augmented reality glasses at the end of last year (Figure 2). This new technology trend has already deserved attention in the world of research [3] and someone has already tried to build his own prototype [4]. On the other hand, someone else has been experiencing wearable computing and home-made augmented reality for the past 20 years [5]. As the market of Apps for smartphones has already had its boom, we suspect another boom is to be expected very soon for wearable-device applications. What is more, interest in first-person perspective was also lately awakened in the field of Cognitive Computation: Froese et al. propose in [6] to address the challenge of investigating first-person experience by designing new humancomputer interfaces, and widely discuss the deep meaning of what they call mind-as-it-could-be from the first-person perspective.

![Microsoft's patent](U.S. Patent Application 20120293548 - November 22, 2012)

Based on these considerations, we see the emergence of a new field of research in computer vision, focused on the users’
point of view. Such a First Person Computer Vision (FPCV) is de facto solicited, and at the same time made possible, by the above mentioned brand-new available technology, namely a wearable computer equipped with a first-person camera framing everyday life. Noticeably, FPV is somehow more specific than simple vision from moving cameras, due to the constraints the device has with the subject and his sub-parts. This specificity might represent an exploitable added value in processing the information gathered, especially for what concerns interactions between the user and the outside world.

As already mentioned, FPV from wearable devices involves information fusion at various complexity and abstraction levels ranging from pure image processing to inference over patterns. We stress here three points and we then start to address the first one throughout the rest of the paper.

- At a low level, image processing must be exploited for object detection, localization, tracking and recognition. Here a the problem of information fusion [7] appears in a manifold of aspects [8] [9] [10].
- Wearable devices such as the above described glasses are going to be equipped with other sensors than a camera. These may include an accelerometer, a gyroscope, a magnetometer, wifi, bluetooth, gps barometer, microphone and many others. Fusing data from this variety of sensor will become an issue in designing applications as for smartphones. Multisensor data fusion is a well known issue and has always drawn the attention of researchers in many fields [11] [12].
- At a higher level of abstraction, scene and behaviour understanding is required for cognitive applications. Here context-based information fusion plays a central role. More in detail, the modelling of the interaction between the user and the outside world often relies on the fusion of "external" and "internal" information (e.g. [13] and [14]).

We start to address the first level of complexity and focus our investigation on hand detection. The reason we focus on hands is exceptionally simple: hands are almost always in our field of view, and are involved in the majority of everyday task (just think about writing, lacing shoes, driving, eating ...). Hands are maybe the principal means we employ to actively interact with the surrounding world, things and persons. So we claim that hands are the best starting point for implementing context-aware functionalities on wearable devices. In addition, new technologies as for instance the Microsoft Kinect and the ever-new Leap Motion controllers [15], seem to be pushing towards hands-free and or even device-free interfaces for an enhanced human/computer interaction. In this context, gesture recognition will play a central role.

The remaining of this work is organized as follows. Section II presents a review of state of the art techniques for hand detection and tracking, together with some consideration on the designing of such algorithms for a wearable device. Section III presents a proposed approach for hand detection and shows preliminary results. Eventually conclusions and additional considerations are drawn in section IV.

II. HAND DETECTION AND TRACKING

Hand detection, tracking and extended tracking (i.e. tracking of hand sub-parts) are problems which have been widely addressed in computer vision. At present, a perfect hand segmentation or accurate hand localization are hardly reached especially under complex conditions. Past approaches are mainly focused on detecting hands from a fixed camera framing a whole person and thus relied on prior knowledge of human silhouettes [16]. The credit for the best degree of accuracy surely goes to depth sensors, which allow for a 3D reconstruction of the scene and a more accurate segmentation of hand shapes [17], thanks to the additional information carried by the depth channel. Yet, we focus on old fashioned digital cameras, as for the moment no such depth sensor is expected to be wearable.

![Hand detection in a human silhouette](image)

Most 2D approaches for hand tracking rely on skin colour features, which looks natural. However, colour features are sensitive to variations in illumination and shadows. A colour correction method is proposed in [18] to overcome this difficulty. It is here claimed that a Gaussian Mixture Model (GMM) can well capture complex variations caused by the difference of human races, gender, age etc. In [19] a Skin HOG model (SHOG) is proposed to construct a robust and efficient hand detector. However, a hybrid approach seems preferable, as also claimed in [20], where the Viola-Jones-like object detection scheme (originally applied to face detection [21]) is combined with a colour based detector, giving satisfactory results for the set of postures considered.

As already pointed out, in the near future more and more videos will be shot by wearable devices, from a first-person point of view. FPV video processing will be more and more requested, raising, among other things, considerable privacy issues. To the best of our knowledge, the problem of detecting and tracking hands from a FPV point of view from a wearable device has never been investigated. We address here some points of this issue and make some considerations.

Good news is that first person perspective can be somehow exploited. Besides, hand tracking is a very peculiar problem. It is common understanding that the more specific a problem is, the more the constraints from the problem itself can be taken advantage of. General issues are often more difficult to be addressed.

- **Number of targets.** Although FPV hand tracking is not a problem with a fixed number of targets, priors on it can be guessed. No more than two targets are allowed, and an equal probability of having zero, one or two targets in the scene can be conjectured.
- **Shape.** Hands from a first person point of view are often framed together with the naked arm. A typical oblong silhouette is shown in Figure 9.
- **Occlusions.** Hand from a first person point of view are hardly occluded. Hand-to-hand or hand-to-object occlusions may of course occur, however hand-to-body occlusions hardly happen.

- **Interactions.** Hands often interact one with the other while performing basic tasks. It has been showed [13] that target interactions can be exploited for improving object tracking.

- **Geometrical constraints.** Hands are linked through arms to the trunk, on which our head is mounted. Degrees of freedom in moving them are then limited by geometrical constraints, which can be then exploited in hand tracking.

- **Personalization** Skin colour features vary a lot from person to person and it is difficult to capture such complex variations, caused by the difference of human races, gender, etc. However, wearable devices are personal (as mobile phones and smartphones), and customized colour models learned from a single user are simpler and more reliable, as it will be shown in III-A.

On the other hand, some bad news comes out in considering first person perspective from a wearable device.

- **Moving camera.** The problem is a typical moving camera issue, which has been widely addressed in literature especially for what concerns moving vehicles. However constant speed or at least certain amount of regularity in the motion must often be hypothesised (see e.g. [22]). Anyway many standard methods such as old fashioned background subtraction for change detection cannot be employed in these circumstances, unless an accurate off-line training phase is performed [23].

- **Framing.** As depicted in Figure 3, hand detection is often performed on well framed ad hoc images. This hardly happens in shooting first person videos, but for specific gestures that could be required for a hypothetical device-free interface.

- **Camera motion estimation.** is complicated by the fact that complex roto-translation matrices are involved, as a head often moves sharply and brusquely. Even integrating data with other sensors such as an accelerometer could be not so easy. The most natural local frame of reference would be the one integral with the device (and thus with the user).

- **Perspective.** The majority of hand detection and gesture recognition methods in literature address the problem from the perspective shown in Figure 3, i.e framing the hands’ palm, from the front. From a first person perspective, the back (or, even worse, the side) of the hand is much more often seen, making gesture recognition harder.

- **Real time requirements.** The video processing algorithm must meet real time requirement, while dealing with the limited computational resources and power supply carried by a wearable device.

### III. Proposed approach

Based on the above considerations we investigate how it could be possible to design a hand tracker, starting from hand detection.

We identify skin colour as the most distinctive and significant feature to be exploited, being also one of the most common used in literature. After concentrating on RGB colour for a while [24], researchers realized that other colour spaces such as L*a*b, HSV [25] and YCbCr [26] proved to be more suitable for colour-based segmentation, not only in the field of skin detection [27]. Various models have been proposed for capturing the information carried by colour, the most common being GMM as already mentioned [18], as Non parametric belief propagation [28] and Viterbi algorithm [29] turned out to be powerful tools for hand tracking.

However, we realized that colour alone does not bring enough information to reliably detect hands, or better, to reliably detect hands only. For this reason we exploit optic flow information in order to filter out false detections, by, roughly speaking, subtracting the global motion of the camera where possible. The way skin-like coloured targets from the background are removed will be explained in details in III-B.

#### A. Colour

Although a GMM better captures complex variations of skin colour due to suntan, gender, age etc. [18] we have already argued that for a single user a single Gaussian is enough to satisfactorily grasp the relevant colour information. The space which better shows clustering of skin pixels turns out to be, from our experience, the CbCr subplane of YCbCr.

The experiment that was carried out is relatively simple and it is depicted in Figures 4, 5 and 6. The device used is a GoPro Hero, outputting a 848x480 video at 50 fps (bitrate is approximately 8000-9000 kbps). Many video sequences were shot, framing a slightly moving hand and gradually changing luminosity in the environment as shown in Figure 4. It can be seen that illumination conditions were stresses to a good extent. Statistics were calculated over the manually drawn red box (the box is drawn in the first frame; it is then checked that it only encompass skin pixels through the video sequence). The typical resulting colour histogram for a generic frame is shown in Figure 5. As one can see very peaked distribution appear for the Cb and Cr channels, while a changing luminosity results in a larger Y histogram. Mean and standard deviation were calculated for each channel for each frames. Means are plotted in Figure 6 for one of the five 1500-frame sequences. Standard deviations are always around 3 (2.95 on average) for the Cb channel and around 4 (4.3 on average) for the Cr channel. It can be seen the extent to which illumination was altered.

Figure 7 shows how the vectors (μ$_{Cb}$, μ$_{Cr}$) cluster in the (Cb, Cr) plane. The covariance matrix of the 2d distribution clearly have eigenvectors which are not parallel to the axis, thus we opted for a bi-dimensional Gaussian to describe the colour model, instead of two separate Gaussians, one for each channel.
Figure 8 shows instead how the scaled feature (Cb/Y,Cr/Y) cluster along what seems to be a straight line (three different hands, three different coefficient). This model does not introduce much improvement, thus we set aside this observation for future works which may include classifying hands sides based on colour.

The model was test on several sequences shot while performing different activities, like drawing, writing on a whiteboard, typing. Figure 9 (a) shows a sample frame. Figure 9 (b) shows colour-based segmentation using a single Gaussian. Given a model \((\mu, \Sigma)\), segmentation is obtain by setting the
condition

\[(x - \mu)^T \Sigma^{-1} (x - \mu) \leq 2, \quad x \in (Cb \otimes Cr) \subset \mathbb{R}^2. \tag{1}\]

It can be clearly seen that objects with skin-like colour (as the mouse pad for instance) are segmented as well as the two hands in the proposed sequence. The way uninteresting targets from the surrounding environment are filtered out is explain below.

\[\text{(a)}\]

![Fig. 9. Colour based segmentation. (a) Sample frame (b) Colour-based segmentation (c) Optic flow-improved segmentation.]

\[\text{(b)}\]

\[\text{(c)}\]

\[\text{Fig. 9. Colour based segmentation. (a) Sample frame (b) Colour-based segmentation (c) Optic flow-improved segmentation.}\]

**B. Optic flow**

We propose that hands doing things hardly move jointly with the head, rather they show different displacements from one frame to another. For this reason we employ optic flow to estimate the average movement of the camera based on the flow vectors calculated through the method proposed in [30] and exploiting the features proposed in [31] and refined in [32]. This way, things moving disjointly from the head can be identified for they show different optic flow vectors associated to their interest points. Results are shown in figure 10. The head is almost still, thus the majority of the flow vectors has very little module and a direction which is opposite to the one towards which the head is (slightly) moving. The other hand the moving hand show marked vectors, which vectorially sum up their and head movements.

Blobs which show no interest points, or which flow is similar to the global one are eliminated as shown in Figure 9 (c). Unfortunately this leads in most frames to the removal of the blob generated by the left hand, which lies still on the table. This suggest that the proposed method works for hands which are acting relevantly only.

The two algorithms do not required massive computational resources, however the 50 fps of the GoPro camera are not supported. On average, frame processing time is around 50 ms.

\[\text{Fig. 10. Optic flow}\]

**IV. Conclusion**

In this work we have presented a global view of the current trend in wearable technologies and their effects in computer vision were examined. We are confident with the fact the in the future the portion of computer vision applied to first person vision will become more and more significant, based on the confidence that wearable technologies will soon have a boom similar to the one smartphones recently had. An overview of the issues arising in FPCV was given, together with considerations on how, on the opposite side, the first-person perspective could be exploited in such a field. The importance of hand detection in this context was widely addressed.

Eventually an approach for hand detection in first person videos was proposed and tested on sequences recorded with a GoPro wearable camera. Such an approach strongly relies on the most natural feature usable for skin detection, namely colour. Fusion with camera movement information, extracted from optic flow vectors allow for filtering undesired detected target which are integral with the environment. Unfortunately this also leads to the removal of inactive hand targets, which could be seen either as a negative or as a positive point.

As a future investigation and prosecution of this work we would like to address the problem of training Haar classifiers [21] for recurring back views of hands and specific hand shapes.

**REFERENCES**


