BACKGROUN SUPPRESSION WITH LOW-RESOLUTION CAMERA IN THE CONTEXT OF MEDICATION INTAKE MONITORING

Gabriel Dauphin
Laboratoire de Traitement et Transport de l'Information
Institut Galilée, Université Paris 13, France

Sami Khanfir
Unité de recherche en Technologies de l’Information et de la Communication à l’ESTT, Tunisie

ABSTRACT

As the aging population is growing, new challenges are arising to provide a safe living environment with remote medical monitoring to allow elderly people to stay at home. This paper is concerned with the monitoring of medication intake. A new technique is proposed for background suppression designed to achieve indoor monitoring for a given video capture device, including low-cost commercially available cameras or webcams with low capturing resolution. The true background image is supposed to be found in the test video sequence, as it is thought to be possible in this application.

The background suppression process can be thought of as a quality measure with reference; the reference being the background image. Instead of taking into account findings on human visual system (HVS), the proposed technique is actually based on measurements of noise output from video capture device.

Experimental results are presented, comparing foreground detection by the proposed technique, two published background suppression algorithms, and three well-known quality measures.

Index Terms—background suppression, medication intake monitoring, noise-robustness, quality measure, multiscale morphology.

1. INTRODUCTION

As the aging population is increasing in the developed countries and as the cost of hospitalization is high, new challenges are emerging to provide a safe environment with remote medical monitoring to allow elders to stay at home. Many dread the idea of living in a rest home and would probably favor using new technologies. In several countries, a number of demonstration projects and demonstration smart houses are being developed and tested [1]. An important research exists on technological devices [2], microphones [3, 4] and/or cameras [5] for activity recognition of fall detection, pointing gesture recognition [6] and medical intake.

This paper is concerned with medication intake monitoring as in [7, 8, 9] and in [10] with a stereo-camera. These different techniques are generally composed of four steps. The first step may be a background suppression, more often this step is a skin segmentation. There may be a preprocessing (lighting correction and gaussian filtering). The second step aims at detecting and tracking some of the following entities: mouth, face, hand, glass (usually opaque) and medication bottle. The third step handles occlusions between these entities. The last step detects activity through analysis of sequences of occlusion. These techniques have been tested on video captures whose foreground is composed of entities that seem to have a resolution ranging from 8 pix/cm to 20 pix/cm. These video captures seem to have little or no visible distortion.

This paper focuses on background suppression which remains a challenging computer vision problem: [11, 12] are robust to gradual and sudden illumination change; [11, 13] are robust to small movements of parts of the background (for instance waving trees); [14] considers a background model image that includes texture elements; [15] is region-based and works with low resolution; [16] has an improved update mechanism for the background image; [17] assumes that the true background image is available.

In Section 2, we present our experimental setup which provides video captures of low resolution and show a hand made ground truth. In Section 3, experimental results of two published algorithms are shown and compared to the ground truth. In section 4, we show that with this experimental setup, background suppression process can be thought of as a quality measure with reference. Our new technique is presented in section 5, it uses morphology operations. Section 6 presents the experimental results. Finally, we conclude this paper in Section 7.

2. THE PROPOSED EXPERIMENTAL SETUP

The experiment takes place in a room with a camera set in a corner near ceiling. Illumination remains unchanged throughout the experiment. This illumination is electric light of low intensity resulting in underexposed video. For the comfort of the reader, all images extracted from these underexposed videos have undergone a gamma correction ($\gamma = 1/2$). In the first part of the video capture there is no actor and the glass is placed elsewhere. In the second part, (the camera has not moved), the actor is sitting on the bed. A pill lies on a white plate and a glass of water is standing nearby. The actor takes the pill, drinks the glass of water and puts the glass back. The true background is available, it is obtained by averaging the frames of the first part of the video capture. The test video showing the actor taking medicine is extracted from the second part of the video capture. Figure 1 is extracted from the test video, it shows the experimental setting. A shadow cast by the actor can be seen in the middle of the image showing the importance of illumination change. A hand-made ground truth is also drawn on this figure. The small blob on the left of the image indicates the disappearance of the glass.

The camera used is a Kodak Easyshare M1063 providing video streams of 480 × 640 pixels at 15 frames per second. Because the camera is far from the scene, objects appear on the video with a reduced number of pixels: 4 pix/cm. Figure 2 is a close-up view of the glass, in which, the individual pixels can easily be observed. Figure 3 is a close-up view of the yellow painted wall, in which fluctuations in colour can be seen. Namely there is a red coloration on the center of the image slightly on the right.
With this specific experimental setup, background image is no longer an output of the background suppression technique, but an input. Background suppression technique can be thought of as pixel-classifier into foreground/background, where classification depends on the colour of a pixel and its differences from the colour of the corresponding pixel in the background image. Is there a difference because of the camera’s noise, or because an entity has appeared on the scene at that specific time and location?

3. RESULTS OF TWO BACKGROUND SUPPRESSION ALGORITHMS FOR COMPARISON

Background suppression techniques described in the literature seem to be unfit for our experimental setup. Two algorithms have been tested on our datasets, they did not provide adequate foreground/background classification.

ViBe, described in [16], operates mainly on a pixel level: background is modeled as a set of colors and foreground/background classification depends on the similarity of the pixel’s colour with a subset of the set of colours. When it is found similar, the background’s model is updated in a random fashion, one of the colours defining the background’s model is replaced by the pixel’s colour. Figure 4 shows the results obtained: only a small part of the true foreground has been detected: part of the face, the left shoulder, parts of the knees and the upper part of the ankles. Part of the left shadow has also been mistakenly detected, showing lack of robustness with respect to change of illumination. By modifying parameters used by ViBe, a larger foreground layer was detected, however it was at the expense of an increase number of noisy pixels detected as foreground. An other attempt was to insert the true background image as a first part of the video, it did not improve the results. The difficulties seem to arise from the very slow movements of the actor and the quick fluctuations of noise.

Fig. 2. Glass-close-up view of figure 1

Fig. 3. Close-up view of figure 1 showing the yellow wall in the upper left corner.

Fig. 4. Black points: foreground layer detected by the ViBe algorithm, [16]. Thick grey/ green curve: hand made ground truth. Only the left lower part of the image is shown here.

An application of gradient field transformation to background suppression is proposed in [17]. The true background is assumed to be available and its edges are suppressed from all frame of the test video, leaving an estimate of the foreground robust to time varying illumination. This method is able to deal with homogeneous regions as it propagates information from edges during the integration of the
modified gradient field.

Fig. 5. Foreground layer recovered with the algorithm described in [17]: bright or dark colours indicate the foreground, greyish colours indicate the background.

Figure 5 shows the results obtained: the foreground layer detected includes the upper part, the lower part and the hands of the actor, and also, a region above the left shoulder and an even larger region of the bedspread. Robustness with respect to illumination did not work properly: the shadow cast by the actor has been replaced by an increased illumination above the left shoulder of the actor, perhaps because a global change of illumination was expected. It is true that the bedspread has moved as the actor sat on the bed, however the algorithm shows a too high sensitivity with respect to such small movements.

In keeping with this viewpoint, we propose to see background suppression process as a quality measure with reference.

4. BACKGROUND SUPPRESSION PROCESS CAN BE THOUGHT OF AS A QUALITY MEASURE WITH REFERENCE

In the last three decades, an important research field concerns the development of quality measures that model how people perceive image and video quality. Evaluation of these quality measures is based on statistical experiments involving a large database of distorted test images/videos and involving a large number of viewers whose opinions on the test images/videos are collected into a Mean Opinion Score (MOS). A good quality measure should correlate well with the MOS. To this purpose many quality measures are based on Human Visual System (HVS).

Quality measures can be classified according to the availability of a reference image/video, with which the distorted image is to be compared. With reference image/video quality means that a complete reference image/video is assumed to be known. Such quality measures are usually implemented in two stages. In the first stage a visibility map is locally evaluated, indicating to what extent the local differences between the test image/video and the reference image/video are visible. In the second stage, all this visibility map is collapsed into a single quality value. No-reference quality measure implies that the reference image/video is not available. Reduced-reference image/video quality measures implies that the reference image/video is only partially available in the form of a set of extracted features. The latters come available as side information to help evaluate the quality of the distorted image. Reviews on image and video quality measures can be found in [18, 19].

With these definitions and as the background image is available, background suppression process can be thought of as a video quality measure with reference, where the reference is the true background image and the visibility map indicates the foreground detected.

By comparing the background image and each frame of the test video, any viewer would be able to figure out, with high precision, the contours of the foreground. Hence quality measures, fully-compliant with HVS, would provide high-quality detection of the foreground. Unfortunately complexity of HVS makes it difficult to design an HVS-inspired quality measure providing, in all circumstances, adequate foreground/background separation. Three quality measures are tested.

The most commonly used quality measure is Peak Signal to Noise Ratio (PSNR), it is expressed in terms of the logarithmic decibel scale and defined as the root mean of the square differences between the grey-level values of pixels in two images. Although not mentioned in literature, a visibility map can be derived from PSNR [18]. It can be defined as absolute value of differences between the two images, which then could be collapsed into the correct single value of PSNR by computing an $L_2$-norm. Figure 6 shows in black/blue the contour of the thresholded PSNR-derived visibility map between the true background and a frame of the test video.

The second quality measure tested is the Structural SIMilarity (SSIM) index [21]. Figure 7 shows the contours of the thresholded visibility map of SSIM between the true background image and a frame of the test video. The detected foreground is the actor, its shadow on the right and the ghost of the glass that has disappeared, see figure 1. Closer look reveals that contours are not correctly positioned. Mathematical and statistical proximity of PSNR and SSIM [22] suggests that a different PSNR-derived map of visibility might yield also interesting results in detecting foreground.
components for each pixel by use of the euclidean norm on colour
three colour components of the differences are collapsed into one
between each frame
higher colour fluctuations are most unlikely to occur.
determined by finding thresholds on flat patches in such a way that
findings the threshold at which a very low-contrast sinusoïdal grat-
HVS-inspired quality measures. Such functions are determined by
are found by conducting a large number of tests. Actually this idea
the corresponding pixel in the background image. Such indications
that some robustness to change of illumination.

The third quality measure tested is the Visible Difference Predictor [20]. Figure 6 shows in light grey/cyan the contours of the
thresholded visibility map of VDP between the true background im-
age and a frame of the test video. The detected foreground is much
precise, it covers the actor and comprises only a small part of its
shadow showing some robustness to change of illumination.

We propose a new technique which is based on the specific cam-
era used in the experimentation.

5. THE PROPOSED TECHNIQUE

We propose a camera-based background suppression technique. It
uses a large set of morphological operations on thresholded images
as in a multiscale morphological approach. Morphological opera-
tions have already been used in the field of background suppres-
sion [12], medication intake monitoring [7] and recognition [23].
Their usage here is motivated by their high spatial resolution regard-
ings have already been used in the field of background suppres-
sion [12], medication intake monitoring [7] and recognition [23].
Their usage here is motivated by their high spatial resolution regard-

The basic idea is that a pixel is labeled as foreground when-
ever an indication is found showing that this pixel and some of its
neighbours could not have a colour that different from the one of
the camera’s noise is
usually assumed to be a white random process [24] whereas HVS is
less of their lack of frequency selectivity. Indeed camera’s noise is
known to exhibit frequency selectivity [18].

The basic idea is that a pixel is labeled as foreground whenever
an indication is found showing that this pixel and some of its
neighbours could not have a colour that different from the one of
the corresponding pixel in the background image. Such indications
are found by conducting a large number of tests. Actually this idea
is quite similar to the use of contrast sensitivity function (CSF) in
HVS-inspired quality measures. Such functions are determined by
finding the threshold at which a very low-contrast sinusoidal grat-
ing stops looking like a uniform image. Our sensitivity function is
determined by finding thresholds on flat patches in such a way that
higher colour fluctuations are most unlikely to occur.

We first focus on the test video and consider the differences
between each frame $I_{mnt}$ and the background image $I_{m}$. The
de three colour components of the differences are collapsed into one
component for each pixel by use of the euclidean norm on colour
components\|. Using colourees may improve the robustness
with respect to change of illumination, but not to noise. Noise is
more evenly distributed in a classical R,G,B space, than in a more
HVS-consistent coloureespace if only because the noise is produced in
the camera electronics and then linearly uncorrelated. The proposed
resulting difference is

\[ D_{mnt} = \| I_{mnt} - I_{m} \| \]  

We propose different kinds of structured elements. Rectangles $B_{kl}$ of size $(k,l) \in \{1..15\} \times \{1..15\}$ are first considered. These rectangles are rotated by $\theta_q \in \{0, \frac{\pi}{10}, \frac{2\pi}{10}, \frac{3\pi}{10}, \frac{4\pi}{10}, \frac{5\pi}{10}, \frac{6\pi}{10}, \frac{7\pi}{10}, \frac{8\pi}{10}, \frac{9\pi}{10}, \frac{10\pi}{10}\}$ and are de-
noted $r_q(B_{kl})$. We extend these structured elements to successive
frames and denote them by $B_{klq}$ and $r_q(B_{klq})$. Note that no spatio-
temporal rotations are considered here. When used in morphological
operations, these structured elements can be thought of as neigh-
bourhoods of pixels. Namely pixels of coordinates $(m_1, n_1, t_1)$ and $(m_2, n_2, t_2)$ will, at some point, be considered together by the
structured element $r_q(B_{kl})$ if they satisfy the following conditions

\[ \begin{cases} 
( (m_2 - m_1) \cos \theta + (n_2 - n_1) \sin \theta \leq k \\
( (m_1 - m_1) \sin \theta + (n_2 - n_1) \cos \theta \leq l \\
|t_1 - t_2| \leq s 
\end{cases} \]  

Precise definitions of morphological operations can be found in [25].
Erosion of set $A$ by structured element $B$, $A \ominus B$, is the set of all
pixels $x$ that remain in $A$ when moved by all associated translations
of set $B$. Dilation of set $A$ by structured element $B$, $A \oplus B$, is
the set of all pixels $y$ that are translations of elements of $A$ by $B$.
Opening of set $A$ by structured element $B$, $A \ominus B$, is achieved by
first eroding set $A$ by $B$, then dilating the resulting set $B$. As for
neighbourhoods, the output of opening by a structured element does
not depend on the location of the center of this structured element.

For each structured elements, we define a threshold $\eta_{klq}$ and a
set of thresholded pixels $T_{klq}$:

\[ T_{klq} = \{ (m, n, t) \mid D_{mnt} \geq \eta_{klq} \} \]  

Threshold values are assumed to be independent of how structured
elements are rotated, and hence $\eta_{klq}$ and $T_{klq}$ do not depend on $q$.
The union of all groups of neighboring pixels that exceed threshold
$\eta_{klq}$ is $T_{klq} \ominus r_q(B_{klq})$. Hence the estimated foreground is defined as

\[ F = \bigcup_{klq} [T_{klq} \ominus r_q(B_{klq})] \]  

Two training videos were used here to learn noise characteris-
tics of the camera and to set thresholds $\eta_{klq}$. During approximately
one minute, two fixed scenes have been shot with the original un-
moved camera. Figure 8 represents a frame of one of the two training
videos. In this figure colour fluctuations can be seen on the bottom
of the white door which is, actually, uniform. Thresholds $\eta_{klq}$ used
in (3) are determined for each structured element $B_{klq}$ as the highest
minimum bound reached by values of $D_{mnt}$ on neighbouring pixels.

\[ \eta_{klq} = \max_{mnt} [D_{mnt} \ominus B_{klq}] \]  

where $D_{mnt} \ominus B_{klq}$ is the flat eroding of single-valued function $D$
by set $B_{klq}$. $D_{mnt} \ominus B_{klq}$ is defined as the minimum $D$-value in
the spatio-temporal region located at $(m, n, t)$ and of the same size
as $B_{klq}$. 

---

**Fig. 7.** Thin dark/magenta curve: contours of the thresholded vis-
ibility map of the Structural SIMilarity (SSIM) index, [21]. Thick
grey/green curve: hand made ground truth.
6. RESULTS

The proposed technique is computationally expensive. It is implemented in Matlab on a 32-bit single core computer running at 2.66GHz with 4GO memory. Setting the threshold values \( \eta_{kls} \) upon the two training videos takes six hours and separating foreground/background takes five minutes per frame which is 5000 times slower than real time.

Figure 9 shows threshold values \( \eta_{kls} \) for different kinds of structured elements as a function of the number of pixels contained in structured elements \( B_{kls} \). The high values of the graph confirm the difficulties highlighted in section 1. Measured in \( D_{mnt} \), the difference between the colour of a pixel and the corresponding pixel in the average image may reach an equivalent of 250 units out of a range of 442 units. As for neighbourhoods of size \( 8 \times 8 \), the minimal difference between the colour of a pixel member of a neighbourhood and the colour of the corresponding pixel may still reach, at some specific time and location, nearly 50 units.

Figure 10 shows the contours of the foreground detected by the proposed technique using threshold values shown in figure 9. It is very similar to the foreground detected by SSIM in figure 7. This detected foreground is incomplete as hair and face remain separated, only part of the arm is found, the disappearance of the glass is missed. It comprises a large part of the shadow. However it is interesting to note that edges with no shadows, (i.e. located on left and on top of the actor) are precisely delineated. The two first observations are due to insufficient extracted information, to use of high thresholds preventing any false foreground detection but allowing false background detection, and to lack of robustness with respect to change of illumination. The third observation may support the use of morphological filters as opposed to linear filters. Indeed proximity of the colour of a pixel with the colour of the corresponding pixel in the true background image avoids classifying as foreground any neighbourhood comprising this pixel.

7. CONCLUSION

The proposed experimental setup involves a low cost camera positioned further away from the person taking medicine and the availability of the true background image. Two background suppression algorithm have failed to produce adequate foreground/background classification, mainly because these algorithm are not designed for that specific experimental setup. Background suppression process can be thought of as a quality measure with reference. Among the three quality measures tested, SSIM provides fairly accurate foreground/background separation but no invariance to change of illumination. The proposed technique is a set of morphological operations on thresholded images. It uses thresholded values computed on two training videos shooting two fixed scenes. At the expense of a large amount of computations it detects a similar foreground with contours better delineated.

Reducing the computation burden is the main challenge. Establishing sufficient conditions for a pixel or a neighbourhood to be classified as background may enable to give an iterative structure to this algorithm and hopefully reduce significantly the average computation burden. Many other improvements may stem from the proposed ideas. Invariance to change of illumination may be constructed by modifying the computation of \( D \) in (1) and by reducing the weight of the luminance component. Reduced false detection of background may be obtained by reducing the chosen thresholds at the expense of an increase in false detection of foreground. Greater precision may be reached by better controlling the spatio-temporal behaviour of colour components.
8. REFERENCES


