Visual Attention-Guided Approach to Monitoring of Medication Dispensing Using Multi-location Feature Saliency Patterns

Roman Palenichka[1], Ahmed Lakhssassi[1] and Myroslav Palenichka[2]

[1]*University of Quebec,Gatineau, Canada*

[2]*Attentive Vision Technologies, Ottawa, Canada*

Abstract

This paper is dedicated to the development of a computer vision-based system for medication (pills and capsules) identification and counting in order to increase the productivity of medication dispensing and maintain its high safety. The algorithmic basis of the system is the attentive vision approach to robust and fast object detection in images. It consists of time-efficient image analysis by a multi-scale visual attention operator to detect feature-point areas located inside the pill and capsule regions. The attention operator combines a spatial saliency filter with a temporal change (novelty) detector in order to robustly detect salient and object-relevant feature points. The medication recognition algorithm involves a set of image descriptors at the feature-point areas called the multi-location feature-saliency pattern, which fully discriminates between different types of medication. The method detects pills and extracts area-based descriptors without any image pre-segmentation procedure due to the proposed multi-scale attention operator.

1. Introduction

The verification process of medication dispensing consists mainly of medication (pills and capsules) recognition and counting. Traditional (i.e., visual and manual) pill identification and counting is a tedious and error-prone process. Sophisticated and automated pill counting systems exist and some are able to handle pill recognition tasks without significant error. Computer vision-based systems have been proposed more recently [1-3]. Unfortunately, the existing fully automated systems typically are not cost-effective for small pharmacies. Moreover, some counting systems are unable to discriminate between different classes of medication pills without errors.

The conventional computer vision approach to pill recognition and counting is the image segmentation of objects versus background with subsequent descriptor extraction and pill recognition [3, 4]. The image with medication is segmented into separated pill regions and the background that represents the pharmacist's tray. Then, a concise description of each region of a medication pill is obtained for decision making processes. The main disadvantage of the conventional methods using segmentation is the instability of this operation in real imaging conditions. Another drawback is the determination of shape descriptors with a low discriminative power for the whole pill region. For example, the mobile-phone recognition system for medication pill verification [3] is based on a heuristic recognition algorithm with the decision making separately by pill size, shape and color characteristics. This complicates the overall process of pill recognition and negatively influences its performance characteristics.

The alternative method to the image pre-segmentation is the feature point (key-point) method, which belongs to the attentive vision approach to fast and robust image analysis [5, 6]. In this approach, the object recognition is based on the descriptor extraction only in the feature point areas, which are detected by a visual attention operator or key-point detector.

There are two major ways of image descriptor extraction: global approach with a single vector of object scalar descriptors [7] and multi-location approach with a set or a relationship graph of local descriptor vectors. In the global approach, usually long vector of scalar descriptors has to contain object-relevant descriptor components, which represents the entire pill region. The second approach involving multiple and short descriptor vectors is preferred in practice because of its high descriptive power (versatility), possibility of relational representation, stability to distortions, and resistance to occlusions. The local descriptor vectors are extracted at the feature points, where objects of interest or image distinctive features are located. The multi-location feature extraction methods based on the image saliency concept can be subdivided into two categories: fix area-based and multi-scale area-based. The disadvantage of existing area-based methods is the necessity of image

segmentation, which is often unstable and time-consuming operation. Since the existing multi-scale operators work well only at lower scale range, they produce some errors in detecting feature points by locating them at ambiguous edge points [7].

The main goal of the presented work was to eliminate or diminish the drawbacks of existing algorithms in the context of medication recognition tasks. Another objective is to develop an algorithmic software basis for an affordable medication verification system suitable for small pharmacy applications and personal assistive usage (e.g., smart phone-based).

The balance of this paper is focused on image descriptor extraction in the feature point areas for robust recognition of medication pills. It is based on the attentive vision approach to image analysis introduced in Section 2. The proposed concept of multi-scale feature points and the corresponding multi-location area descriptors called the multi-location feature-saliency pattern is an important contribution of this paper (Section 3). Section 4 presents the proposed algorithm for pill recognition, which is based on the matching of descriptor vector sets. Preliminary experimental results are described in Section 5 and conclusions are given in Section 6.

2. The attentive vision approach

The proposed approach to medication recognition and pill counting is inspired by the attention-focusing mechanisms of the human visual system, which serve to time-efficiently locate and reliably recognize objects of interest in complex, changing, and distractive environment [8]. It consists in attention-guided image analysis by detecting attention points and analyzing image areas around these points. Object-relevant attention points represent the feature points which concisely describe the objects of interest such as medication pills. Decision making is usually made by comparing the feature-point area descriptors with the reference ones.

The proposed algorithm of attention-guided image analysis is also a model-based approach to descriptor extraction founded on the *salient disk model* (SDM) of object-relevant image areas [9]. It is a representation of objects of interest as a set of high color-contrast homogeneous areas called salient image disks. The salient disk model ensures the feature-point areas to be robustly detectable, locally unique, and positioned inside the objects of interest.

The attentive vision method exploits two basic spatiotemporal characteristics of the objects of interest in images for the determination of feature-point locations: spatial saliency and temporal change or novelty. Spatial saliency of an object area is proportional to a local contrast value of the image intensity (color) located inside that area with respect to the background intensity or color. Spatial saliency of an area centered at the current analysis point is the necessary condition for the object detection ability. The temporal change in the present application conditions is the appearance of new objects of interest such as the medication pills, which can be called a novelty occurrence.

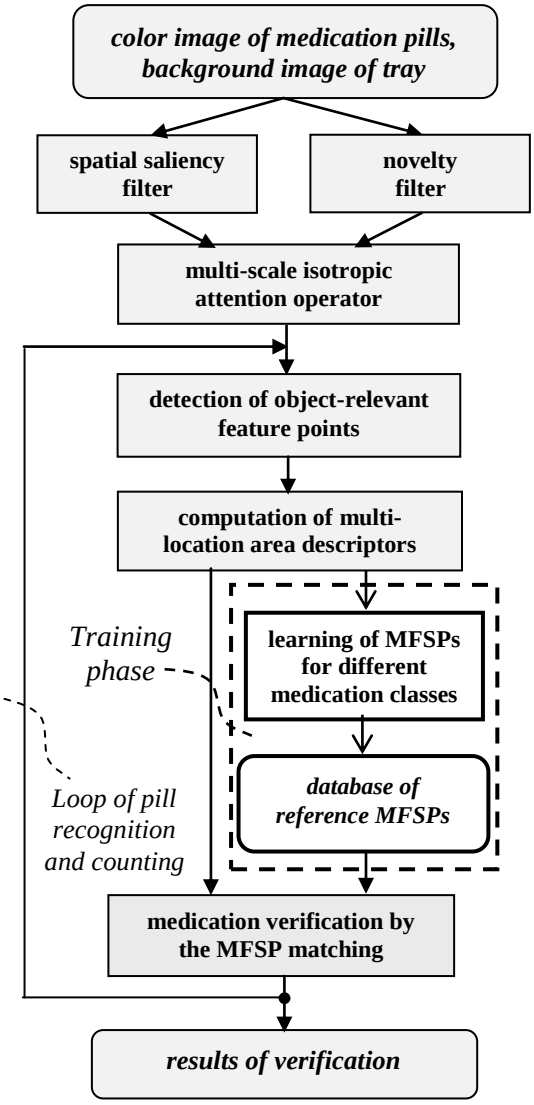

Fig. 1. Global flowchart of the attentive vision approach.

The global flowchart of the attentive vision approach to medication recognition and counting is shown in Fig. 1. It consists of two hierarchical image analysis phases for object recognition: focal and peripheral. These two phases represent the visual attention mechanisms for fast and reliable image analysis. The peripheral analysis phase carries out the entire image plane analysis for rapid object localization to determine the locations of the medication pills. It is mostly implemented through the time-efficient computation of an attention operator, which local maxima indicate the feature-point areas. The focal or foreground image analysis is aimed at the detailed analysis of selected image areas with the spatial saliency and scene novelty, which are located in a neighborhood of a selected feature point. The selected feature point, which neighborhood is being under detailed image analysis is called the focus-of-attention (FoA) point. Local image descriptors are estimated in each feature point area in the neighborhood of the FOA point and united into an image local representation entity called multi-location feature saliency pattern. Pill recognition is implemented by the matching of feature-point descriptors with the reference pill descriptors, which are determined at the learning stage and include all the medication pill classes (Section 4).

3. Detection of feature points and descriptor extraction

The image descriptor extraction is the crucial step in robust attention-guided recognition of medication pills as long as the object recognition proceeds without image segmentation. Misdetection or false detection of attention (feature) points with their descriptors can result in significant recognition errors.

The approach of multiple feature points consists of two coherent steps of descriptor extraction: 1) detection of object-relevant feature points; 2) estimation of transformation-invariant descriptor components for each feature-point area. The detection is implemented as a sequential search for local maxima of a multi-scale attention operator. In order to insure robust detection, the multi-scale attention operator is composed of two parts – spatial saliency filter and novelty (temporal change) filter – which are aggregated into a single function of image coordinates (i,j) and local scale ρ [10]:

$$F[i,j,\rho] = s(i,j,\rho) + \beta \cdot e(i,j,\rho), \qquad (1)$$

where $s(i,j,\rho)$ and $e(i,j,\rho)$ are the spatial saliency and novelty filter values at $(i,,j)$, and β is the novelty coefficient (β>0). The derivation of the attention operator $F[i,j,\rho]$ and optimized value of the coefficient β in the maximum-likelihood sense can be obtained assuming appropriate distributions of two terms in Eq. (1) as random independent variables at the condition of a feature point occurrence at (i,j) [10].

We have used the contrast-based approach to regional (area) saliency definition [10, 11]. The first term $s(i,j,\rho)$ in Eq. (1) is the difference between the local isotropic contrast $c(i,j,\rho)$ and the area inhomogeneity (e.g., intensity local variance) $d(i,j,\rho)$:

$$s(i,j,\rho) = c(i,j,\rho) - d(i,j,\rho). \qquad (2)$$

The local isotropic contrast $c(i,j,\rho)$ at the ρth scale is estimated as the mean absolute deviation in the ring $Q_\rho(i,j)$ with respect to the mean intensity inside the disk region $S_\rho(i,j)$ (Fig. 2). The area inhomogeneity term provides the relative value of the contrast with respect to the area inhomogeneity in order to obtain more robust detection.

The novelty filter in Eq. (1) is the background subtraction operator, which computes the difference $h(i,j)$ between the current image and the background image (i.e., image without medication pills) with subsequent spatial integration. The spatial integration is based on the spatial contrast value of the novelty filter:

$$e(i,j,r) = h_S(i,j,r) - h_Q(i,j,r), \qquad (3)$$

where, r is the diameter of the disk $S_r(i,j)$, $h_S(i,j,r)$ and $h_Q(i,j,r)$ are the mean values of $h(i,j)$ in the disk $S_r(i,j)$ and ring $Q_r(i,j)$, respectively (Fig. 2). Similarly to the spatial saliency term, the temporal change is estimated relatively to the change occurred in the surrounding neighborhood.

In order to discriminate good feature points from all other attention points, each candidate feature point has to be tested for its saliency level, local uniqueness and object relevance [10]. In general, it is implemented as a simple thresholding operation, which involves an adaptive threshold value [5, 10]. For example, the object relevance of a feature point is determined by comparing the novelty filter value with a threshold. In this way, only feature points, which are located inside the pill regions, will be extracted and analyzed in detail.

Four different types of area independent descriptors are extracted to form the local descriptor vector (LDV), which represents the current disk area: 1) planar pose (two coordinates of feature point, local scale and local direction); 2) area shape; 4) color components; 3) surface intensity. The use of these image descriptors is justified

by the salient disk model-based approach [9]. All descriptor components in the LDV are normalized to be contained in the range [-1; 1].

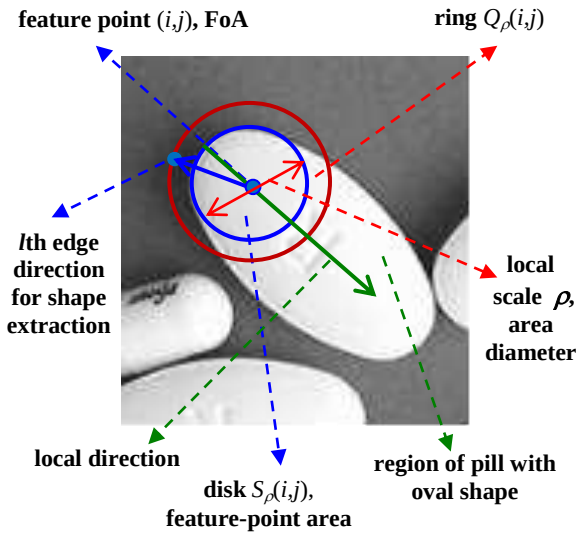

Fig. 2. Example of a feature-point area detected in an oval pill region.

A set of planar shape descriptors – called radial shape pattern – was proposed to achieve robustness of shape descriptors and maintain their simplicity [9]. This algorithm consists of two basic steps: a) determination of local shape direction; and b) estimation of L directional edge descriptors. The local direction as one of the area planar pose descriptors is used at the step (a). The second step consists of directional derivatives at L radial edge points lying in the ring $Q_k(i,j)$ (Fig. 2).

The color descriptors are important elements of a concise description since medication pills can have different colours while their geometry descriptors are the same. Mean values over the disk $S_\rho(i,j)$ for the three color components in the transformed color space have been used in the experiments.

The model-based approach of intensity surface approximation was adopted in order to obtain descriptors of spatial distribution for color intensities in a given region such as the feature-point area [12]. The surface intensity descriptors include the mean value of intensity and three components of the intensity gradient, which are computed within the disk of diameter ρ. The first two gradient components are estimated within smaller sub-areas in two orthogonal orientations, along and across the local direction to achieve the rotation invariance.

One feature point area with its LDV very often cannot be sufficient to describe all the shape, color, and surface varieties of medication pills. Therefore, several feature points with their LDVs are aggregated into a multi-location feature saliency pattern (MFSP) to better discriminate between different pills in real conditions of medication dispensing. The transformation-invariance of the LDV pose descriptors is achieved through their relational values with respect to the central feature point descriptors as the FoA point. In particular, the relational positions of feature point areas in a MFSP are an important element of the MFSP concept in the transformation-invariant object recognition.

The main computational burden of feature-point detection and descriptor extraction proposed in this paper is carried out by the multi-scale attention operator in Eq. (1). The time-efficient implementation of the attention operator is based on the fast recursive algorithms for the computation of local moments of image intensity [13]. They provide fast computation of local spatial saliency and novelty filters independently of the filter window size, which corresponds to the local scale.

4. Medication verification

The medication verification consists in the recognition and counting of medication pills placed on a pharmacist's tray after their preliminary selection by a pharmacist. It is composed of two stages of image analysis: training (i.e., machine learning) and recognition (Fig. 1). The training stage is executed only once as a machine learning procedure, in which images of new medication pills are presented as one pill per image to obtain their representation by image descriptors in the form of a MFSP. The attentive vision approach to object recognition provides a simplified machine learning procedure as well. In the considered application, only few feature point areas were sufficient to identify medication pills by their MFSPs. First of all, the values of *optimized parameters* of the algorithm in Fig. 1, such as the scale range, coefficients, and threshold values are estimated by a simple averaging operation and range determination using the training sample of pill images. For example, the scale range is determined by obtaining the minimum and maximum values of the local size for all the pills available in the training sample.

The maximal number of LDVs in a MFSP is selected during the training stage to discriminate between different medication classes. The learning procedure starts from a single LDV extracted from the pill image of a current class. It corresponds to the feature

point area centered at the global maximum of the multi-scale attention operator. If the single LDV of the FoA-point area is insufficient to discriminate medication pills of different classes then another LDV is added to the current MFSP. This process is continued by adding more feature points with their LDVs until a full discrimination capability is achieved.

The recognition stage basically is the matching of a current MSFP with the reference ones stored in a database of MFSPs for all medication pills. A point-set distance can perform such a matching, which is a distance between two sets of LDVs as points in a multi-dimensional metric space of descriptor components [14]. Since the point sets can have different sizes, the distance between them can be determined by a particular matching algorithm such as the Hausdorff distance, which was also used for general image matching. Other, more complex point-set distances can be applied depending on the description matching task at hand [14].

For this particular application case, we propose a different matching algorithm in order to reduce the computation time needed for the estimation of a point-set distance. It consists in organizing a fast hierarchical search in the database of pill MFSPs for the most similar MFSPs to the current one. The search starts from finding the most similar feature-point areas to the descriptors of the FoA point, which is the most salient LDV of the observed MFSP. The reference MFSPs, which match the FoA point area, provide the most probable locations of other feature points of a pill region. Therefore, only limited locations are checked in the first step for feature point locations by comparing their pre-computed saliency value with a threshold. If a feature point is detected, then its area descriptors are estimated and a distance between the observed LDV and the reference one is computed. This process is continued until all the LDVs are computed and compared with the reference ones. Finally, the medication class is determined by the minimal point-set distance among them if several candidates are found in the database of reference MFSPs. In our experiments (Section 5), the mean value of Euclidean distances between matching LDV pairs has been used as a point-set distance between two MFSPs.

Pill counting operation proceeds together with the recognition of each pill by considering the next most salient feature point as the FoA point (Fig. 1). The above described matching of MFSPs is applied at each step of selecting the next feature point.

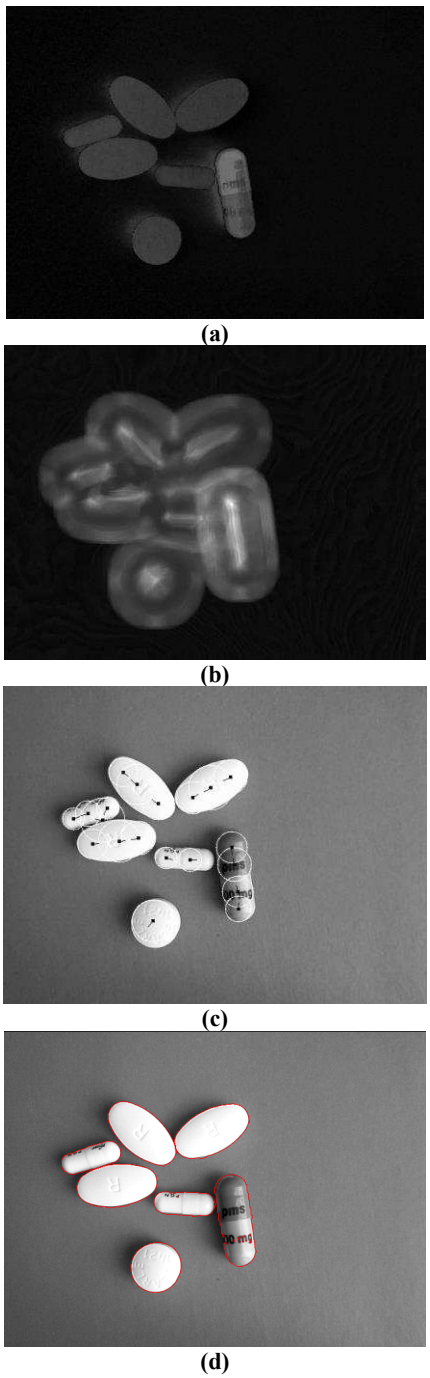

Fig. 3. Extraction of feature points and multi-location area descriptors for pill recognition: novelty map (a); multi-scale attention operator (b); feature point areas with pose descriptors (c); medication pills correctly recognized by the proposed algorithm (d).

5. Experimental results and discussion

The goal of the experiments was the performance evaluation of the automated medication verification by the visual attention-guided approach. It consisted of two major evaluation tasks: 1) accuracy of feature point detection and descriptor estimation; 2) accuracy of medication recognition.

The task of feature point detection is crucial in the overall process of medication recognition and counting since failure do detect a feature point often results in and undetected pill or capsule, which contains this feature point. Another problem is connected with the separate location of pills, which can often be adjacent by their edges and therefore incorrectly recognized or miscounted.

The medication recognition algorithm proposed in this paper uses image descriptors derived from the salient disk model (SDM) representation of feature-point areas [9]. Since the descriptor set was significantly modified with regards to the original SDM-based algorithm [8], the accuracy of feature point extraction and descriptor estimation was evaluated in comparison with the original algorithm [9] as well as the SIFT method [5] (Table I). The error value in Table I is the normalized absolute difference between the current descriptor estimate and its reference value as the ground truth. Semi-synthetic, computer graphic generated images have been used in the experiments in order to provide the ground truth values of descriptor components. Different perturbations of intensity (color) are applied to the semi-synthetic image in order to imitate the real imaging conditions. The images were rotated or scaled and then transformed back in order to investigate the transformation invariance of the descriptor extraction. The error of shape descriptor estimates was averaged over all the 16 components of the radial shape pattern (Section 3).

In the evaluation of comparative accuracy of pose descriptors (two coordinates), we have applied the attention operator DoG (Difference of Gaussians), which was used in the original SIFT algorithm to extract feature points [5]. This operator is less sensitive for feature-point areas with low color (intensity) contrast because of using the Gaussian averaging over the same central area. Moreover, it does not take into account the pill area homogeneity, which contributes to the accuracy of feature point detection. In the proposed isotropic attention operator, the area local contrast is computed in Eq. 2 relatively to the area homogeneity that increases the contrast value at the center of homogeneous low-contrast areas.

The pill recognition algorithm used in these experiments is composed of three major steps: 1) detection of feature points; 2) extraction of area descriptors for each detected feature point; 3) MFSP-based matching of feature point areas. Fig. 3 illustrates the proposed visual attention-guided algorithm for pill recognition with its intermediate steps: background subtraction map (a), multi-scale isotropic attention operator (b) feature point areas with pose descriptors and medication pills shown with the red contours. In the image 3c, the centers of feature-point areas are indicated by black dots and local directions are shown as black line segments. The local scale value at the feature points corresponds to the diameter of the white circle. In these experiments, the attention operator uses novelty coefficient β=0.3 and scale range R=11. Fig. 3b shows that the local maxima of the attention operator are located at the centers or medial lines of the medication pills and capsules. The MFSP was composed of two LDVs only, which correspond to two most salient feature points. Each LDV is composed of 40 descriptor components, including pose, shape, color and surface intensity. The local shape is represented by the radial shape pattern, which contains 16 directional edge descriptor components (Section 3). It should be noted that the contour extraction (red contours in Fig. 3d) of the pills and capsules recognized by the proposed algorithm has been carried out exclusively for the visualization purpose. The proposed recognition algorithm is not relying on any image pre-segmentation procedure.

The performance of medication (pills and capsules) recognition was tested on test samples of pills and capsules of different classes placed together in different combinations. Some examples are shown in Fig. 4.

Table I: Accuracy of descriptor extraction.

Algorithm	Location of feature point	Local direction	Shape descriptors	Mean color intensity
Original SDM-based, intensity perturbations	0.06	0.10	0.12	0.07
SIFT, intensity perturbations	0.13	Not available	0.15	0.11
Modified SDM-based, intensity perturbations	0.08	0.09	0.10	0.05

Fig. 4. Some examples of test color images for pill recognition and counting experiments**.**

Since the first performance test evaluates the accuracy of feature point detection, the second test mostly will indicate the descriptor extraction performance if the feature point detection works without major errors. The step of MFSP-based matching has no significant influence on the performance since it uses a simple distance-based comparison of a current MFSP with the reference ones.

The medication recognition results obtained in the conducted experiments are given in Table II as the recognition performance standard measures: precision and recall. The total number of medication test images (with different combinations of medication pills) used in these experiments was 162, while the total number of different classes of medication pills was 23. The recognition tests have been carried out for two types of multi-scale attention operators: 1) Spatial saliency + Novelty; 2) Spatial saliency only. The test result shows that it operates very well even without using the novelty filter in the multi-scale attention operator. That extends its application scope and reduces the runtime significantly. For comparison, the recognition experiments have been implemented with the segmentation-based algorithm similar to the one proposed in [3].

The proposed algorithm of feature-point extraction and MFSP-based pill identification was also tested on its computational efficiency. The efficiency of the attentive vision approach using MFSP descriptors is mostly determined by the time efficiency of the feature point detection because descriptor extraction proceeds only in a limited number of feature point areas. Besides, it exploits intermediate results of the attention operator computation. The computational complexity of the multi-scale attention operator is $O(1)$ per pixel and per scale since it does not depend on the operator window (scale) size due to the involvement of fast recursive algorithms for local moments computation [13]. The proposed algorithm of descriptor extraction for pill recognition has shown the speedup of 2.7 with respect to the corresponding SIFT-based algorithm using the DoG multi-scale attention operator [5]. The conditions (parameter values) of the comparative efficiency tests were the same as the conditions of the accuracy tests described above.

Table II: Performance of medication recognition.

Method for identification of medication pills	Algorithm for pill (feature point) detection	Precision rate	Recall rate
MFSP	Spatial saliency + Novelty	0.96	0.92
MFSP	Spatial saliency	0.93	0.91
Global regional descriptors	Color image segmentation	0.92	0.89

6. Conclusions

A visual attention-guided algorithm for the medication dispensing verification using feature-point descriptors of color images is proposed. It is based on the pill image representation as a MFSP and has the following advantageous characteristics. *First*, the robustness of descriptor extraction and formation of MFSPs is achieved through the selection of feature points only in stable, locally unique and object-relevant image locations. *Second*, descriptors are made invariant to eventual similarity transformations and affine changes of intensity. *Third*, the method extracts multi-scale area-based descriptors without image pre-segmentation due to the proposed multi-scale attention operator. *Fourth*, the medication recognition is organized as a time-efficient algorithm for feature-point detection and subsequent matching of the current MFSP with the reference ones using a point set distance.

References

[1] R. N. Hamilton, *US Patent* 6738723 - Pharmacy pill counting vision system, 2004.

[2] J. R. Wootton, V. V. Reznack, and G. Hobson. *US Patent* 6535637 - Pharmaceutical pill recognition and verification system, 2003.

[3] A. Hartl, "Computer-vision based pharmaceutical pill recognition on mobile phones", *Proceedings of CESCG 2010*: *14th Central European Seminar on Computer Graphics*, pp. 51-58, 2010.

[4] T. Evgeniou, *et al*., "Image representations and feature selection for multimedia database search", *IEEE Trans. KDE*, 15(4), pp. 911-920, 2003.

[5] D. G. Lowe, "Distinctive image features from scale-invariant key-points", *Int. J. of Computer Vision*, 60(2), pp. 91-110, 2004.

[6] K. Mikolajczyk and C. Schmid, "Performance evaluation of local descriptors", *IEEE Trans. PAMI*, 27(10), pp. 1615-1630, 2005.

[7] K. Mikolajczyk *et al*., "A comparison of affine region detectors ", *Int. J. of Computer Vision*, 65(1), pp. 43–72, 2005.

[8] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans. PAMI*, Vol. 20(11), pp. 1254-1259, 1998.

[9] R. Palenichka, M. Petrou, Y. Kompatsiaris, and A. Lakhssassi, "Model-based extraction of area descriptors using a visual attention operator", *Proc. Int. Conference ICPR 2012, Tokyo*, pp. 853-857, 2012.

[10] R. Palenichka, A. Lakhssassi and M. Zaremba, "A spatiotemporal attention operator using isotropic contrast and regional homogeneity", *Journal of Electronic Imaging*, 20(2), pp. 023018 (2011): 1-15, 2011.

[11] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," *Proc. IEEE Conference CVPR 2012*, pp. 733-740, 2012.

[12] K. Mikolajczyk *et al*., "A comparison of affine region detectors ", *Int. J. of Computer Vision*, 65(1), pp. 43–72, 2005.

[13] V. Di Gesù and R.M. Palenichka, "Fast recursive computation of local axial moments", *Signal Processing*, 81(2), pp. 265-273, 2001.

[14] T. Eiter and H. Mannila, "Distance measures for point sets and their computation", *Acta Informatica*, 34(2), pp. 109-133, 1997.