HUMAN ACTION POSELETS ESTIMATION VIA COLOR G-SURF IN STILL IMAGES

M. Favorskaya *, D. Novikov, Y. Savitskaya

Institute of Informatics and Telecommunications, Siberian State Aerospace University, 31 Krasnoyarsky Rabochy av., Krasnoyarsk, 660014 Russian Federation - (favorskaya, novikov_dms)@sibsau.ru, juliklapa@yandex.ru

Commission WG V/5, WG III/3

KEY WORDS: Human Action, Poselets, Gauge-SURF, Random Forest, Still Image

ABSTRACT:

Human activity is a persistent subject of interest in the last decade. On the one hand, video sequences provide a huge volume of motion information in order to recognize the human active actions. On the other hand, the spatial information about static human poses is valuable for human action recognition. Poselets were introduced as latent variables representing a configuration for mutual locations of body parts and allowing different views of description. In current research, some modifications of Speeded-Up Robust Features (SURF) invariant to affine geometrical transforms and illumination changes were tested. First, a grid of rectangles is imposed on object of interest in a still image. Second, sparse descriptor based on Gauge-SURF (G-SURF) invariant to color/lighting changes is constructed for each rectangle separately. A common Spatial POselet Descriptor (SPOD) aggregates the SPODs of rectangles with following random forest classification in order to receive fast classification results. The proposed approach was tested on samples from PASCAL Visual Object Classes (VOC) Dataset and Challenge 2010 providing accuracy 61-68% for all possible 3D poses locations and 82-86% for front poses locations regarding to nine action categories.

1. INTRODUCTION

Recognition of human active actions is employed in many tasks of computer vision, e.g., video surveillance, human detection, activity analysis, scene analysis, image annotation, image and video retrieval, augmented reality, human-computer interaction, among others. Conventional methods of human action recognition use motion information in videos (Laptev and Lindeberg T, 2003, Zhen et al., 2013), which can be applied for object capturing and then for trajectory motion analysis as the temporal component in human action classification. During action classification, the spatial component plays a significant role that makes reasonable to develop methods for human action recognition in still images.

Methods used to estimate the human poses can be classified in four categories: model-based, example-based, pictorial structure-based, and poselet-based approaches. The model-based methods predefine a parametric body model and find the pose matching based on labelled extracted features. A graphical model of human-object interactions was developed by Gupta et al. (Gupta et al., 2009) including reach motions, manipulation motions, and object reactions. Such models are often built on silhouette-based representation of body parts or edge information. The main disadvantage deals with difficulties in design of parametric body model and pose sub-models.

The example-based methods do not use a global modelling structure and store a set of images with corresponding pose descriptions. In this framework, two problems appear as relevant descriptors and fast search. Three shape descriptors – Fourier descriptors, shape contexts, and Hu moments were compared by Poppe and Poel (Poppe and Poel, 2006) for representation of human silhouettes. They experimented with deformed silhouettes robustness to body sizes, viewpoint, and noise. Shakhnarovich et al (Shakhnarovich et al., 2003) proposed interesting hashing-based search technique for pose estimation relevant to pose examples in a large database. A fast pruning method based on shape contexts in order to speed up the search for similar body poses was presented by Wang et al. (Wang et al., 2006). High computational cost of searches in the high-dimensional spaces inside large datasets is the major drawback of example-based methods.

The pictorial structure-based methods represent poses as the cues using prior information of a human body structure (Felzenszwalb and Huttenlocher, 2005). In this approach, the histogram-based methods prevail. It may be circular histograms of spatial and orientation binning (Ikizler et al., 2008) or the most popular Histogram of Oriented Gradients (HOG) (Dalal, and Triggs, 2005) with multiple modifications. The last research was the pioneer investigation in pose descriptor construction based on non-negative matrix factorization. The action classes were represented by the HOGs of pose primitives with following simple histogram comparison for action recognition. This technique works well in typical cases but it fails in occlusions or significant changes of camera viewpoints. To overcome these problems, Delaitre et al. (Delaitre et al., 2010) proposed a Bag-of-Features (BoF) approach for human action recognition in still images in combination with Support Vector Machine (SVM) classification. They combined the statistical and part-based representations integrating a person-centric description in cluttered background.

The poselet-based methods provide rich information about locations of body parts. They are built on 3D pose images. The poselets were introduced by Bourdev and Malik (Bourdev and Malik, 2009) for person detection in natural framework. The detailed literature review of methods from the fourth category is

* Corresponding author

presented in following section. Better pose descriptor means a wide invariance to various geometric transforms, lighting, viewpoints, and image warping in general case. This difficult task is in the field of current and following investigations. Our contribution deals with the proposed color Gauge-SURF with selection of special imposed grids in human body image for human action recognition. The original Gauge-SURF was extended by invariance to color and lighting changes.

In the following, Section 2 gives a brief review of high-level and low-level evaluations of poselets. The G-SURF background is presented in Section 3. The proposed methodology of poselets estimation is explained in Section 4. The experimental results are presented and analyzed in Section 5. The paper is concluded in Section 6.

2. RELATED WORK

The poselet as a subject of interest is 2D still non-segmented image of configured body parts in 3D space (head, shoulders, arms, torso, legs) capturing a part of neighbouring background. Thus, the poselet of single pose is a collection of 2D still images received from various shooting viewpoints, in different scales and lighting conditions. Sometimes, the presence of articulation makes the pose estimation harder. Although the pose estimation is used for human action recognition, one may consider the pose estimation as a separate task of computer vision. Figure 1 depicts some examples of human actions.

Figure 1. Human pose examples

Since only the spatial information is available in still images, one can represent information as high-level cues and low-level cues/features. Description of human body, body parts, action-related objects, human object interaction, and scene context are included in high-level cues. Typical low-level features are a Dense sampling of Scale Invariant Feature Transform (DSIFT), HOG, Shape Context (SC), GIST, and some other features. Short surveys of high-level and low-level evaluation methods are situated in Sections 2.1 and 2.2, respectively, while existing 3D-based poselet methods are discussed in Section 2.3.

2.1 High-level Evaluation

A human body image is an important cue in human action recognition, which can be detected automatically or manually labelled. Usually a bounding box is used to indicate a location of a person. Some approaches extract features in areas within or surrounding the human bounding boxes. Delaitre et al. (Delaitre et al., 2010) defined a person setting in each image in one and a half time more that the sizes of human bounding box. Then these regions are resized up to 300 pixels (in larger size) and analyzed using low-level features.

Some methods extract contour information of human body and body parts from still images. Wang et al. (Wang et al., 2006) exploited the overall coarse shape of human body as a collection of edge points obtained via Canny edge detector. Then the received features were classified and categorized into different actions. Also semantic features can be used to describe the actions in images with the human body (Yao et al., 2011). The attributes were related to verbs in a human language and remained visual words from annotation system.

Yang et al. (Yang et al., 2010) developed a coarse example-based poselets representation, when each body part may have more than 20 poselets concerning to different body poses. They constructed a set of four corresponding body parts $L = \{l_0, l_1, ..., l_{k-1}\}$ denoting the upper-body, legs, left-arm, and right-arm. Often a graph is a good model to represent the relations between different body parts. Raja et al. (Raja et al., 2011) constructed a graphical model containing six nodes: the action label and five body parts correspond to head *H*, right-hand *RH*, left-hand *LH*, right-foot *RF*, and left-foot *LF*. The links between nodes encode action-dependent constraints on the relative positions of body parts.

In many scenarios, a person relates with other objects, e.g. phone, ball, animal, etc. These objects serve as a source of reliable information about category of human action. Some approaches analyse individual objects separately while other methods consider them as a scene context. As a result, some methods were developed as a Human Object Interaction (HOI), for example, weakly supervised method proposed by Prest et al. (Prest et al., 2012).

2.2 Low-level Evaluation

The high-level evaluation is usually based on various low-level scores. The DSIFT features are extracted from many image patches with following clustering to obtain a limited number of "keywords", which are grouped in a codebook. Many methods use the DSIFT features due to their possibility for direct classification of human actions. One can mention the researches of Delaitre et al. (Delaitre et al., 2010), Yao et al. (Yao et al., 2011), etc.

The HOG descriptor is very popular for pedestrian detection (Dalal and Triggs, 2005). The HOG descriptor counts the occurrences of discrete gradient orientations within a local image patch similar to the edge orientation histogram, the SIFT, and the SC. The SC is useful to detect and segment the human contour; however, this technique is crucial for high-level cue representation of human body silhouettes.

The spatial envelop or GIST was proposed by Oliva and Torralba (Oliva and Torralba, 2001). A set of spatial properties in a scene can be computed by the GIST method, which provides the abstract category representations of a scene based on integrated background information. This approach has been used by Gupta et al. (Gupta et al., 2009), Prest et al. (Prest et al., 2012), among others. One can mention the development of other approaches based on SURF (Bay et al., 2008), Circular

Histogram of Oriented Rectangles (CHORs) (Ikizler et al., 2008), Adaboost classifiers (Gupta et al., 2009).

2.3 Towards to 3D Representation

All variety of poselet-based methods may be classified according to various criteria. A viewpoint dependence/ independence in 2D/2D spaces, respectively, can be considered the main criterion. Bourdev and Malik (Bourdev and Malik, 2009) were the first, who formulated the task of human pose estimation and recognition as 3D object representation. They constructed the body part detectors trained from annotated data of joint locations of people and based on patches similarities. As a result, the poselet activation vector consisting of poselets inside the bounding box was introduced. The SVM classifier was used to recognize these patches. The distribution of these joints and personal bounding boxes can be obtained to each poselet.

Russakovsky et al. (Russakovsky et al., 2012) developed an Object-Centric spatial Pooling (OCP) approach for detection an object of interest. The local OCP information is used to pool the foreground and background features. Khan et al. (Khan et al., 2013) applied a comprehensive evaluation of color descriptors with combination of shape features.

Some methods deal with detection of view-independent objects using 3D object models. Glasner et al. (Glasner et al., 2011) used a viewpoint estimation method for rigid 3D objects from 2D images by voting method for efficient accumulation of evidence. This method was tested on rigid car data. Fidler et al. (Fidler et al., 2012) developed a method for localizing objects in 3D space by enclosing them within tightly oriented 3D bounding boxes. This model represents an object class as a deformable 3D cuboid by anchors of body parts in 3D box. Hejrati and Ramanan (Hejrati and Ramanan, 2012) developed a two-stage model, when, first, a large number of effective views and shapes are modelled using a small number of local view-based templates and, second, these estimates are refined by an explicit 3D model of the shape and viewpoint.

Another issue is a computational speed required for view-independent object detection because the most methods generate the classifiers at multiple locations and scales. Sometimes, the shared features are extracted to reduce a number of classifiers and the runtime complexity correspondingly. Razavi et al. (Razavi et al., 2010) used an extension of Hough-based object detection and built a shared codebook by jointly considering several viewpoints. Tosato et al. (Tosato et al., 2010) modelled a human image as a hierarchy of fixed overlapping parts. Each part was trained using a boosted classifier learned using Logicboost algorithm. Velaldi et al. (Velaldi et al., 2009) introduced a three-stage SVM classifier combining linear, quasi-linear, and non-linear kernels for object detection.

3. G-SURF BACKGROUND

The original SURF defines the determinant of the approximate Hessian matrix in the points, where the determinant has maximum values (Bay et al., 2008). The Hessian matrix $\mathbf{H}(\mathbf{p}; \sigma)$ is defined by Equation 1:

$$\mathbf{H}(\mathbf{p};\sigma)=\begin{bmatrix} L_{xx}(\mathbf{p},\sigma) & L_{xy}(\mathbf{p},\sigma) \\ L_{xy}(\mathbf{p},\sigma) & L_{yy}(\mathbf{p},\sigma) \end{bmatrix} \quad (1)$$

where $\mathbf{p} = (x, y)^T$ = a point in an image I

σ = a scale factor

$L_{xx}(\mathbf{p}, \sigma)$ = a convolution of an image $I(\mathbf{p})$ in a point $\mathbf{p}$ with a Gaussian second order derivative along direction OX

A convolution $L_{xx}(\mathbf{p}, \sigma)$ is determined by Equation 2:

$$L_{xx}(\mathbf{p},\sigma) = I(\mathbf{p}) * \frac{\partial^2}{\partial x^2} g(\sigma) \quad (2)$$

and similarly for $L_{xy}(\mathbf{p}, \sigma)$ and $L_{yy}(\mathbf{p}, \sigma)$ along diagonal and OY directions, respectively.

Alcantarilla et al. (Alcantarilla et al., 2013) developed a novel family of multi-scale local feature descriptors called as Gauge-SURF (G-SURF). In this case, every pixel in the image is described by 2D local structure. The multi-scale gauge derivatives are invariant to rotations and shifts. Additionally, they describe the non-linear diffusion processes. The use of G-SURF makes blurring locally adaptive to the region so that noise becomes blurred, whereas details or edges remain unaffected. Such local structures are described by Equation 3:

$$\begin{aligned} \vec{w} &= \left(\frac{\partial L(\mathbf{p},\sigma)}{\partial x}, \frac{\partial L(\mathbf{p},\sigma)}{\partial y} \right) = \\ &\frac{1}{\sqrt{L_x^2(\mathbf{p},\sigma) + L_y^2(\mathbf{p},\sigma)}} \cdot \left(L_x(\mathbf{p},\sigma), L_y(\mathbf{p},\sigma) \right); \\ \vec{v} &= \left(\frac{\partial L(\mathbf{p},\sigma)}{\partial y}, -\frac{\partial L(\mathbf{p},\sigma)}{\partial x} \right) = \\ &\frac{1}{\sqrt{L_x^2(\mathbf{p},\sigma) + L_y^2(\mathbf{p},\sigma)}} \cdot \left(L_y(\mathbf{p},\sigma), -L_x(\mathbf{p},\sigma) \right) \end{aligned} \quad (3)$$

where $\vec{w}$ = a gradient vector

$\vec{v}$ = a perpendicular direction vector

σ = a kernel's standard deviation or scale parameter

$L(\mathbf{p}, \sigma)$ = a convolution of image $I(\mathbf{p})$ with 2D Gaussian kernel $g(\mathbf{p}, \sigma)$

Using gauge coordinates, one can obtain a set of derivatives invariant to any order and scale. The second-order gauge derivatives $L_{vv}(\mathbf{p}, \sigma)$ and $L_{ww}(\mathbf{p}, \sigma)$ are in special interest. They can be obtained as a product of gradients in $\vec{w}$ and $\vec{v}$ directions and the 2×2 second order derivatives or Hessian matrix provided by Equation 4:

$$\begin{aligned} L_{ww}(\mathbf{p},\sigma) &= \frac{1}{L_x^2(\mathbf{p},\sigma) + L_y^2(\mathbf{p},\sigma)} \begin{pmatrix} L_x(\mathbf{p},\sigma) & L_y(\mathbf{p},\sigma) \end{pmatrix} \\ &\begin{pmatrix} L_{xx}(\mathbf{p},\sigma) & L_{xy}(\mathbf{p},\sigma) \\ L_{yx}(\mathbf{p},\sigma) & L_{yy}(\mathbf{p},\sigma) \end{pmatrix} \begin{pmatrix} L_x(\mathbf{p},\sigma) \\ L_y(\mathbf{p},\sigma) \end{pmatrix}; \\ L_{vv}(\mathbf{p},\sigma) &= \frac{1}{L_x^2(\mathbf{p},\sigma) + L_y^2(\mathbf{p},\sigma)} \begin{pmatrix} L_y(\mathbf{p},\sigma) & -L_x(\mathbf{p},\sigma) \end{pmatrix} \\ &\begin{pmatrix} L_{xx}(\mathbf{p},\sigma) & L_{xy}(\mathbf{p},\sigma) \\ L_{yx}(\mathbf{p},\sigma) & L_{yy}(\mathbf{p},\sigma) \end{pmatrix} \begin{pmatrix} L_y(\mathbf{p},\sigma) \\ -L_x(\mathbf{p},\sigma) \end{pmatrix} \end{aligned} \quad (4)$$

The G-SURF descriptor is based on the original SURF descriptor. Mention some modifications such as Modified Up-right SURF (MU-SURF) descriptor based on Haar wavelets

responses and two Gaussian weighting steps, the Center Surround Extremas (CenSurE) for approximation the bi-level Laplacian of Gaussian using boxes and octagons, the Speeded Up Surround Extrema (SUSurE), which is a fast modification of the MU-SURF, and the CenSurE descriptors for mobile devices. These modifications well describe edges, angles, boundaries of unknown objects in an image. However, they poorly concern color information, which is useful in analysis of still images.

4. ESTIMATION OF POSELETS

The poselets estimation is concerned to identification task, when a number of classes is restricted by a finite set of human poses describing in feature space by some descriptors. The complexity deals with another issue – the great variants of 3D images of a single pose mapping in 2D still images. Nowadays, this problem has not been solved, and many authors develop heuristic algorithms, more or less successful.

The structure of bounding boxes, in common case, grid, for poselet detection is discussed in Section 4.1. The proposed color G-SURF family is represented in Section 4.2. Section 4.3 provides a classification procedure based on random forest in order to identify a testing poselet.

4.1 Imposed Grid for Poselet Detection

For poselet recognition, it is required to capture a salient part of one's pose from a given viewpoint and impose a set of corresponding rectangle boxes at given orientation, position, and scale. Many authors use the predetermined aspect ratios sometimes with normalization of distance between hips and shoulders. Bourdev and Malik (Bourdev and Malik, 2009) used the poselets with following of aspect ratios 96 × 64, 64 × 64, 64 × 96, and 128 × 64 pixels. Their algorithm was trained using 300 poselets of each type of pose. Then a model predicting a bounding box for each poselet was fitted. Additionally, the overlapping bounding boxes were considered with overlapping area more 20%. These scores were added to the basic set. The 1200 dimensional vector to estimate the human pose was constructed, besides the poses of head and torso were considered separately. This approach was developed by Ko et al. (Ko et al., 2015), when five aspect ratios such as 96 × 64, 64 × 64, 64 × 96, 64 × 128, and 128 × 64 pixels were used in order to consider variations of human poses. These authors modified the selecting algorithm of the action poselets using the modified Hausdorff distance with Epanechnikiv kernel. The descriptor is based on Oriented Center-Symmetric Local Binary Patterns (OCS-LBPs) due to their low computational complexity. Such approach proposes the overlapping of aspect ratios with preliminary rough body parts segmentation. Background is involved in this grid. Often segmentation is implemented manually.

It is mindless to use the predefined aspect ratios as well as the random aspect ratios. It will be reasonable to find Regions Of Interest (ROIs) or pose primitives accurately. If a video sequence is available, then this issue is the solved problem, when the ROIs are detected from the temporal domain in previous frames (Favorskaya, 2012; Favorskaya et al., 2015). Also the hybrid methods based on optical flow and learning of salient regions are possible (e.g., Eweiwi et al. 2015).

If a single still image is available, then one of pixel-based segmentation methods can extract large areas with identical texture under assumption that such large areas are the body parts. These areas representing the bounding boxes are the basis of the SPOD constructions, which later are aggregated in a common SPOD for classification. This task is similar to automatic image annotation. Our recommendations deal with the use of J-SEG algorithm or similar ones. If this assumption is not fulfilled, then additional human segmentation methods are required in order to receive the cropped image of human body.

Let us suppose that cropped still image of human body is obtained, and six aspect ratios are used including head, torso, left arm, right arm, left leg, and right leg. Each of the arm and leg boxes is composed from two sub-boxes including hands and feet. A number of imposed aspect ratios can be reduced due to human action and human position. Moreover, a number of imposed aspect ratios serve as a weak classifier for human action categorization. For example, the active action "phoning" may include an image of head and any arm while the active action "running" may involve images of torso and both legs.

4.2 The Proposed Color G-SURF Family

Under various conditions of color/lighting changes in human pose images, it is important to develop a descriptor invariant to these changes. Our contribution deals with the development of original G-SURF, which is invariant to geometrical (affine) distortions. Family of the proposed color G-SURF descriptors includes the following components.

The rg G-SURF (rgG-SURF) includes the chromaticity components r and g invariant to scale and light changes. The rg histogram is based on the normalized RGB color model, where the components r and g describe the color information by Equation 5 (b is redundant as $r + g + b = 1$):

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \dfrac{R}{R+G+B} \\ \dfrac{G}{R+G+B} \\ \dfrac{B}{R+G+B} \end{pmatrix} \tag{5}$$

where R, G, B = Red, Green, Blue components in RGB color space, respectively

r, g, b = normalized components

Because of normalization, the components r and g are scale-invariant and invariant to lighting changes, shadows and shading (Gevers et al., 2006). This descriptor has 45 dimensions.

The Opponent G-SURF (OppG-SURF) analyzes three channels in the opponent color space using G-SURF descriptor (Equation 6):

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \dfrac{R-G}{\sqrt{2}} \\ \dfrac{R+G-2B}{\sqrt{6}} \\ \dfrac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{6}$$

where O_1, O_2 = two channels in the opponent color space providing color invariance

O_3 = the channel in the opponent color space providing intensity invariance

The Hue G-SURF (HueG-SURF) is constructed by concatenation of a hue histogram in Hue Saturation Value (HSV) color space with a histogram of G-SURF descriptors. Such descriptor is scale-invariant and shift-invariant with respect to light intensity due to a Hue histogram. In a Hue histogram, the hue is weighted by the saturation of a pixel and reflected the instabilities in hue. This histogram has 36 dimensions.

The transformed RGB color G-SURF (RGBG-SURF) descriptor is computed for each normalized RGB channel by Equation 7:

$$\begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} = \begin{pmatrix} \frac{R-\mu_R}{\sigma_R} \\ \frac{G-\mu_G}{\sigma_G} \\ \frac{B-\mu_B}{\sigma_B} \end{pmatrix} \quad (7)$$

where μ_R, μ_G, μ_B = means of the distribution in RGB channels computed in a chosen region of image, respectively

σ_R, σ_G, σ_B = standard deviations of the distribution in RGB channels computed in a chosen region of image, respectively

The RGBG-SURF is invariant to scale and shift with respect to light intensity (Van de Sande et al., 2009), while a classic RGB histogram is not invariant to changes in lighting conditions The histogram has 45 dimensions.

Also one can mention additional color descriptors with some invariance to color/lighting conditions such as the color moment histogram, the hue-saturation descriptor, the color names, the discriminative color descriptor, among others. However, the experiments show that the rg G-SURF, the OppG-SURF, the HueG-SURF, and the RGBG-SURF provide better results in poselets estimation.

4.3 Classification Procedure

Various classifiers have been applied for object classification such as SVM, Boosting-algorithms, and random forest. The SVM classifier is the well proved technique for general classification. However, SVM is not suitable, when the features have high dimensionality and a set of analysed images is huge. The Boosting-algorithms such as AdaBoost or GentlBoost are the popular machine learning methods but their performance depends critically on the choice of weak classifiers.

A random forest is one of the most popular tree-based classification approaches, which is effective in a large variety of high-dimensional tasks such as object detection and object tracking (Ko et al., 2013). The random forest is an ensemble classifier of several randomized decision trees, which have a capacity to analyze big data at high training and runtime speeds. Each tree is grown using some type of randomization. The structure of each tree is binary, and all trees are created in a top-down manner (Breiman, 2001). During the training stage, the random forest uses a random subset from training data initially. In common case, at node n the training data D_n are split interactively into left and right subsets using a threshold and the split function according to Equation 8:

$$\begin{aligned} D_l &= \left\{ i \in D_n \middle| F(v_i) < th \right\}; \\ D_r &= D_n \setminus D_l \end{aligned} \quad (8)$$

where D_l, D_r = left and right subsets, respectively
th = a threshold
$F(v_i)$ = a split function
v_i = ith feature vector

The threshold T is selected randomly in the range $T \in (\min F(v_i), \max F(v_i))$. The use of an ensemble of trees trained with small random subsets increases a speed of training and reduces amount of overfitting. Because a random forest dismisses the spatial information of local regions (patches) within a detection window, it produces some false positives, especially when a background has a similar appearance.

In this study, the common SPOD in a view of the concatenated histograms (OX and OY) of color G-SURF descriptors in each bounding box is used as a feature for random forest classification. In training stage, each tree T is constructed based on a set of examples Ex_i of a poselet $Ex_i = (e_i, c_i)$, where e_i is a poselet example and c_i is a class label of a poselet example. The positive poselet examples are marked by a class label $c_i = 1$, and the negative poselet examples gain a class label $c_i = 0$. Samples of the other body parts including a background are concerned to the negative poselet examples. During tree construction, each leaf node L stores the class information C_L of the examples. When only positive poselet examples are reached at node N, then $C_L = 1$. Value of C_L is proportional to a number of positive poselet examples. A split function is assigned to each non-leaf node in such manner that the uncertainties in class labels ought to be reduced towards leaves. In testing stage, the current element descriptor of a poselet is checked using the created binary trees by sequential comparison with value of split function in each node.

5. EXPERIMENTS AND DISCUSSIONS

The PASCAL Visual Object Classes (VOC) Dataset and Challenge 2010[1] was used in training and testing stages. It includes 10,103 and 9,637 images for training and testing, and also 23,374 and 22,992 objects for training and testing, respectively. For experiments, a part of dataset called Action Classification Taster Challenge 2010 was applied. In this part, each person image with 2D pose was annotated by "layouts" of the person with bounding boxes: head, hands, and feets. In total, 322 images containing 439 people were annotated as the training/validation set. The test set involves 441 images with a single person. Figure 2 depicts example images for nine action categories (phoning, walking, running, taking photo, playing instrument, riding bike, riding horse, reading, and using computer) with detected G-SURFs in bounding boxes. Sometimes, a non-significant number of SURFs and G-SURFs can be detected in bounding box due to very small sizes of box, e.g., hand and foot boxes. In these cases, a Kuwahara filter with radius $R = 3$ was applied in order to receive increased number of feature points. As one can see in Figure 3, where the scaled fragments of images are represented, this additional procedure helps to decide this problem partly.

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

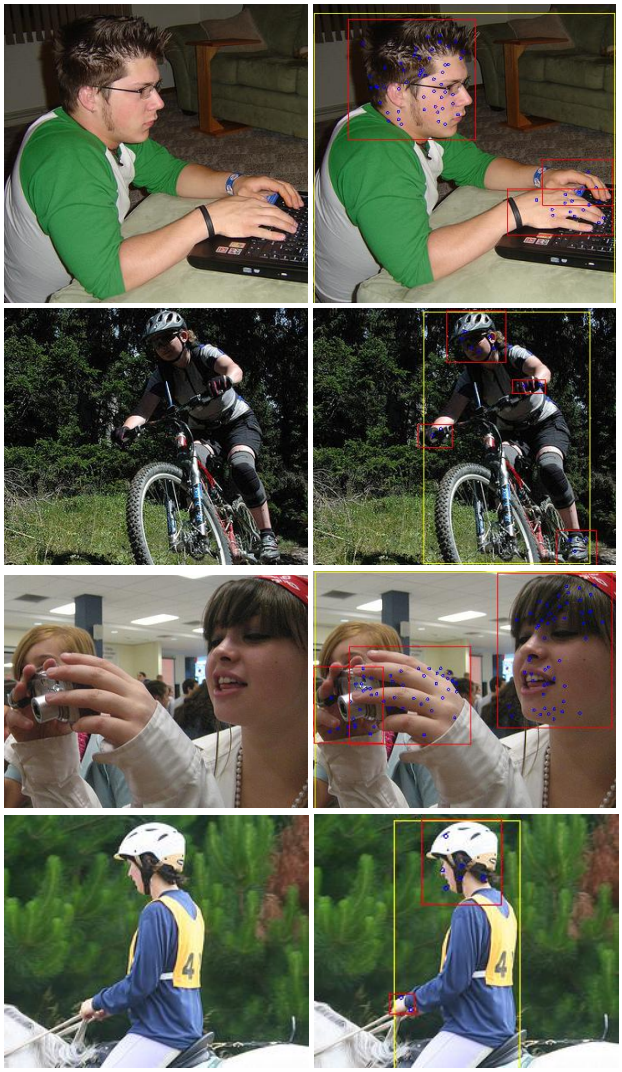

Figure 2. The processed person layouts with detected G-SURF

Figure 3. Scaled fragments with increased number of G-SURFs

Then histogram descriptors had been built in bounding boxes concerning to various body parts. They were classified by random forest technique. During training stage, the separate random forest for heads, hands, and feet were constructed. An example of random forest training (one from a set of trees) is located in Figure 4.

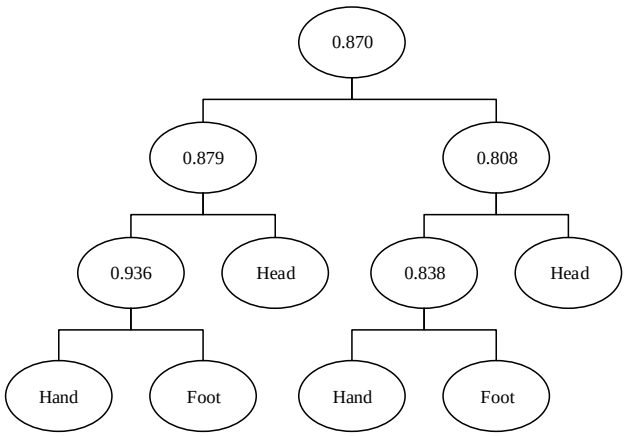

Figure 4. Random forest training

Except SURF and G-SURF, all four types of color G-SURF were calculated in corresponding color spaces. The experimental results show that the rgG-SURF and the HueG-SURG demonstrated the worst results of recognition and false ratios because they are not robust to light intensity scaling and the G-SURF provided a partial robustness. Table 1 includes the precision evaluation for nine action categories.

Category	SURF	G-SURF	OppG-SURF	RGBG-SURF
Phoning	73.7	75.1	76.8	76.2
Walking	62.2	63.5	64.6	64.1
Running	59.6	60.4	61.3	61.0
Taking photo	77.8	78.1	78.9	78.4
Playing instrument	61.1	62.3	63.1	62.9
Riding bike	58.3	59.7	60.2	59.7
Riding horse	76.6	78.2	78.4	78.2
Reading	68.2	69.7	70.3	69.3
Using computer	83.4	85.1	86.0	85.4

Table 1. Precision results for nine action categories (%)

The analysis of received results shows the common tendency: more sizes of bounding boxes provide large number of feature points that leads to better recognition results. Thus, phoning, playing photo, riding horse, and using computer are characterized with better precision results.

Tables 2 and 3 contain false rejection rates and false acceptance rates. Also one can see that errors have high values. This may be explained by restriction of current statement of problem. In future, additional procedures, e.g., a graph of body parts, skin detection, skeleton representation, Kinect data analysis, and analysis of surrounding objects, among others, will permit the promising results against current results. The current research shows that the analysis of bounding boxes limits estimators significantly.

Category	SURF	G-SURF	OppG-SURF	RGBG-SURF
Phoning	38.9	37.5	36.1	35.9
Walking	40.1	39.2	38.6	38.3
Running	40.2	38.1	37.2	37.1
Taking photo	39.0	38.3	36.8	36.4
Playing instrument	49.1	47.8	45.4	45.1
Riding bike	50.3	49.6	47.9	47.7
Riding horse	51.4	50.2	48.7	48.4
Reading	34.8	33.7	32.5	32.3
Using computer	37.2	36.4	35.2	35.0

Table 2. False rejection rates for nine action categories (%)

Category	SURF	G-SURF	OppG-SURF	RGBG-SURF
Phoning	28.2	27.3	25.6	25.2
Walking	25.1	24.4	23.1	22.8
Running	28.2	27.1	25.7	25.3
Taking photo	21.8	21.2	20.2	20.1
Playing instrument	49.1	47.9	46.0	45.6
Riding bike	38.9	37.4	36.3	35.8
Riding horse	34.4	33.2	32.5	32.1
Reading	29.8	28.5	27.3	26.9
Using computer	27.2	26.1	25.2	24.8

Table 3. False acceptance rates for nine action categories (%)

False rejection rates are higher in comparison to false acceptance rates. It can be explained by usual difficulties in object recognition using still images relative to video-based object recognition in the spatio-temporal domain.

6. CONCLUSION

In this study, the poselets estimation is proposed for recognition of human active actions. Our approach deals with accurate segmentation of body parts in a still image by a possibility of the temporal data extraction from a video sequence or pixel-based segmentation method in order to receive the images of body parts. The experiments with G-SURF lead to the descriptor invariant to color/lighting conditions. The SPOD is based on color G-SURF family. Better results were received using OppG-SURF and RGBG-SURF. For classification of common SPOD, a random forest classification was used as a fast and effective procedure during work with big data representing various poselets of various human active actions. PASCAL VOC Dataset and Challenge 2010 provided the test material for the training and the testing stages. The precision for nine action categories such as phoning, walking, running, taking photo, playing instrument, riding bike, riding horse, reading, and using computer achieves on the average of 61-68% for all 3D poses locations and 82-86% for front poses locations regarding to action categories.

REFERENCES

Alcantarilla, P.F., Bergasa, L.M., Davison, A.J., 2013. Gauge-SURF descriptors. *Image and Vision Computing*, 31(1), pp. 103-116.

Bay, H., Tuytelaars, T., Van Gool, L., 2008. Surf: Speeded up robust features. *J. Computer Vision and Image Understanding*, 110(3), pp. 346-359.

Bourdev, L., Malik, J., 2009. Poselets: body part detectors trained using 3D human pose annotations. In: *The IEEE Conference on Computer Vision*, Kyoto, Japan, pp. 1365-1372.

Breiman, L., 2001. Random forests. *Mach. Learn*. 45(1), pp. 5-32.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: *The IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, pp. 886-893.

Delaitre, V., Laptev, I., Sivic, J., 2010. Recognizing human action in still images: a study of bag-of-features and partial-based representations. In: *The British Machine Vision Conference*, Aberystwyth, UK, pp. 1-11.

Eweiwi, A., Cheema, M.S., Bauckhage, C., 2015. Action recognition in still images by learning spatial interest regions from videos. *Pattern Recognition Letters*, 51, pp. 8-15.

Favorskaya, M., 2012. Motion estimation for objects analysis and detection in videos. In: Kountchev, R., Nakamatsu, K. (Eds) *Advances in Reasoning-Based Image Processing Intelligent Systems*, ISRL vol. 29, Springer-Verlag Berlin Heidelberg, pp. 211-253.

Favorskaya, M., Pyankov, D., Popov, A., 2015. Accurate motion estimation based on moment invariants and high order statistics for frames interpolation in stereo vision. In: Tweedale, J.W., Jain, L.C., Watada, J., Howlett, R.J. (Eds) *Knowledge-Based Information Systems in Practice*, SIST vol. 30, Springer International Publishing Switzerland, pp. 329-351.

Felzenszwalb, P.F., Huttenlocher, D.P., 2005. Pictorial structures for object recognition. *Int. J. Comput. Vis.*, 61(1), pp. 55-79.

Fidler, S., Dickinson, S., Urtasun, R., 2012. 3D object detection and viewpoint estimation with a deformable 3D cuboid model. In: *The Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, pp .620-628.

Gevers, T., van de Weijer, J., Stokman, H., 2006. Color feature detection: an overview. In Lukac, R., Plataniotis, K.N. (Eds) *Color image processing: methods and applications*. University of Toronto, Ontario, Canada: CRC press, pp. 203-226.

Glasner, D., Galun, M., Alpert, S., Basri, R., Shakhnarovich, G., 2011. Viewpoint-aware object detection and pose estimation., In: *The IEEE Conference on Computer Vision*, Barcelona, Spain, pp.1275-1282.

Gupta, A., Kembhavi, A., Davis, L.S., 2009. Observing human-object interactions: using spatial and functional compatibility for recognition. *IEEE Trans. on Pattern Anal. Mach. Intell*., 31(10), pp. 1775-1789.

Hejrati, M., Ramanan, D., 2012. Analyzing 3D objects in cluttered images, In: *The Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, pp.602-610.

Ikizler, N., Cinbis, R.G., Pehlivan, S., Duygulu P., 2008. Recognizing actions from still images. In: *The International Conference of Pattern Recognition*, Tampa, Florida, USA, pp. 1-4.

Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A.D., Lopez A.M., Felsberg M. 2013. Coloring Action Recognition in Still Images. *Int. J. of Computer Vision*, 105(3), pp. 205-221.

Ko, B.C., Kwak, J.Y., Nam, J.Y., 2013. Human tracking in thermal images using adaptive particle filters with online random forest learning. *Opt. Eng*. 52(11), pp. 1-14.

Ko, B.C., Hong, J.H., Nam, J.Y., 2015. View-independent object detection using shared local features. *J. of Visual Languages and Computing*, 28, pp. 56-70.

Laptev, I., Lindeberg, T., 2003. Space-time interest points. In: *The IEEE International Conference on Computer Vision*, Nice, France, pp.432-439.

Oliva A., Torralba A., 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis*., 42(3), pp. 145-175.

Poppe, R., Poel, M., 2006. Comparison of silhouette shape descriptors for example-based human pose recovery. In: *The International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, pp.541-546.

Prest, A., Schmid, C., Ferrari, V., 2012. Weakly supervised learning of interactions between humans and objects. *IEEE Trans. on Pattern Anal. Mach. Intell*., 34(3), pp. 601-614.

Raja, K., Laptev, I., Pérez, P., Oisel, L., 2011. Joint pose estimation and action recognition in image graphs. In: *The IEEE International Conference on Image Processing*, Brussels, Belgium, pp.25-28.

Razavi, N., Gall, J., Gool, L.V. ,2010. Back projection revisited: scalable multi-view object detection and similarity metrics for detections. In: *The European Conference on Computer Vision*, Heraklion, Crete, Greece, pp.620-633.

Russakovsky O., Lin Y., Yu K., Fei-Fei L., 2012. Object-centric spatial pooling for image classification, In: *The European Conference on Computer Vision*, Florence, Italy, pp.1-15.

Shakhnarovich, G., Viola, P., Darrell, R., 2003. Fast pose estimation with parameter-sensitive hashing. In: *The IEEE International Conference on Computer Vision*, Nice, France, pp.750-757.

Tosato, D., Farenzena, M., Cristani, M., Murino, V., 2010. Part-based human detection on Riemannian manifolds. In: *The IEEE International Conference on Image Processing*, Hong Kong, China, pp. 3469-3472.

Van de Sande, K.E.A., Gevers, T., Snoek, C.G.M., 2009. Evaluating color descriptors for object and scene recognition. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 32(9), pp. 1582-1596.

Velaldi, A., Gulshan, V., Varma, M., Zisserman, A., 2009. Multiple kernels for object detection. In: *The IEEE Conference on Computer Vision*, Kyoto, Japan, pp.606-613.

Wang, Y., Jiang, H., Drew, M.S., Li, Z.N., Mori, G., 2006. Unsupervised discovery of action classes. In: *The IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, pp. 1654-1661.

Yang, W., Wang, Y., Mori, G., 2010. Recognizing human actions from still images with latent poses. In: *The IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, pp. 2030-2037.

Yao, B., Jiang, X., Khosla, A., Lin, A., Guibas, L., Fei-Fei, L., 2011. Human action recognition by learning bases of action attributes and parts. In: *The IEEE International Conference on Computer Vision*, Barcelona, Spain, pp.1331-1338.

Zhen, X., Shao, L., Tao, D., Li, X., 2013. Embedding motion and structure features for action recognition. *IEEE Trans. on Circuits Syst. Video Technol*., 23(7), pp. 1182-1190.