Spatial-Temporal Granularity-Tunable Gradients Partition (STGGP) Descriptors for Human Detection

Yazhou Liu, Shiguang Shan, Xilin Chen, Janne Heikkila,
Wen Gao, and Matti Pietikainen

Key Laboratory of Intelligent Information Processing, Institute of Computing Technology
Chinese Academy of Sciences (CAS), China
Machine Vision Group, Department of Electrical and Information Engineering
University of Oulu, Finland
{yzliu,sgshan,xlchen,wgao}@jdl.ac.cn
{Janne.Heikkila,Matti.Pietikainen}@ee.oulu.fi

Abstract. This paper presents a novel descriptor for human detection in video sequence. It is referred to as spatial-temporal granularity -tunable gradients partition (STGGP), which is an extension of granularity-tunable gradients partition (GGP) from the still image domain to the spatial-temporal domain. Specifically, the moving human body is considered as a 3-dimensional entity in the spatial-temporal domain. Then in 3D Hough space, we define the generalized plane as a primitive to parse the structure of this 3D entity. The advantage of the generalized plane is that it can tolerate imperfect planes with certain level of uncertainty in rotation and translation. The robustness to the uncertainty is controlled quantitatively by the granularity parameters defined explicitly in the generalized plane. This property endows the STGGP descriptors versatile ability to represent both the deterministic structures and the statistical summarizations of the object. Moreover, the STGGP descriptor encodes much heterogeneous information such as the gradients' strength, position, and distribution, as well as their temporal motion to enrich its representation ability. We evaluate the STGGP on human detection in sequence on the public datasets and very promising results have been achieved.

1 Introduction

Human detection research has received more and more attention in recent years because of increasing demands in practical applications, such as smart surveillance system, on-board driving assistance system and content based image/video management system. Even through remarkable progress has been achieved [1,2,3,4,5,6,7], finding the human is still considered as one of the hardest task for object detection. The difficulties come from the articulation of human body, the inconsistency of clothes, the variation of the illumination and the unpredictability of the occlusion.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part I, LNCS 6311, pp. 327–340, 2010.

Human detection from the still images has been one of the most active research fields during the recent years. Varieties of features have been invented to overcome the difficulties mentioned above. Earlier works for human detection started from Haar-like features, which have been applied to face detection task successfully [8,9,10]. Because of the large variation of human clothes and background, some researchers turned to the contour based descriptors. Gavrila [11] presented a contour based hierarchical chamfer matching detector. Lin et al. [12,13] extended this work by decomposing the global shape models into parts to construct a parts template based hierarchical tree. Ferrari el al. [14] used the network of contour segments to represent the shape of the object. Wu and Nevatia [15] used edgelet to represent the local silhouette of the human.

After the invention of the SIFT descriptor[16], more researchers have used the statistical summarization of the gradients to represent human body. Such as the position-orientation histogram features proposed by Mikolajczyk et al. [17]; the histograms of oriented gradients (HOG) proposed by Dalal et al. [18,19] and it's improvements [20]; the covariance matrix descriptor proposed by Tuzel et al. [21]; and the HOG-LBP descriptor proposed by Wang et el. [2]. More recently, granularity-tunable gradients partition (GGP) for human detection was proposed by Liu et al. [22], in which granularity is used to define the spatial and angular uncertainty of the line segments in the Hough space. By adjusting the granularity, GGP provides a container of descriptors from deterministic to statistic.

Even with these powerful representation methods, the appearance of human body is still not discriminative enough, especially in some complex environments. Therefore, some works use the motion information to improve the performance of human detection. As mentioned in [23], certain kinds of movement are characteristics of humans, so detector performance can potentially be improved by including motion information. Viola et al. [24] used the Haar-like filters to extract the appearance from the single image and extract the motion information from the consecutive frames. By including the motion information, they can improve the performance of their system remarkably. Dalal et al. [23] used orientated histograms of differential optical flow to capture the motion information of the human, and then they combined the motion descriptors with histogram of oriented gradient appearance descriptors. The combined detector can reduce the false alarm rate by a factor of 10. Similar improvements have also been reported by Wojek et al. in their resent work [25].

These works show that incorporating the motion and appearance information is a promising way to improve the performance of human detection. Therefore, this work extends the granularity-tunable gradients partition (GGP) [22] from the image domain to the spatial-temporal domain. This new descriptor is referred to as spatial-temporal granularity-tunable gradients partition, or STGGP for short. In STGGP, human and their motions are modeled in the joint spatial-temporal domain. The spatial-temporal volume representations has been widely used in the action recognition research, as in [26,27,28], and very promising results have been reported. But for human detection research, most of the well-

Fig. 1. The spatial-temporal representation of human body

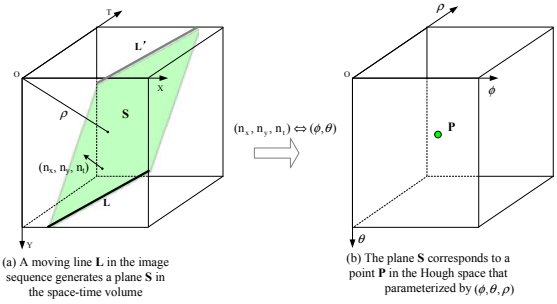
known methods, such as[24,23,25], model the human appearance and motion information as two separate channels. This work considers the moving human body as a 3-dimensional entity in the spatial-temporal domain. Then we use the *generalized planes* to parse the structure of this 3D entity.

The *generalized planes* are defined in the Hough space, which extend the representation of the plane from a point to a cuboid region. The size of the cuboid region is related to a certain level of robustness to rotation and translation uncertainty. Therefore, by changing the size of the cuboid region, the robustness can be controlled explicitly. Hence, a family of descriptors with different representation abilities can be generated that range from the specific geometrical representation to the statistical summarization of the object. This multiple representation property is referred to as *granularity-tunability* and the size parameters of the cuboid region is referred to as *granularity parameters*, or granularity for short. This property enables the STGGP descriptor to represent the complex human pattern in the spatial-temporal domain.

The rest of the paper is organized as follows: Section 2 introduces the human representation method in the spatial-temporal domain; Section 3 defines the generalized plane; Section 4 presents the mapping method of the generalized plane from the Hough space to the spatial-temporal domain; Section 5 gives the computational details; and Section 6 contains the experimental results.

2 Spatial-Temporal Volume Representation of Human Body in Video

The spatial-temporal volume (STV) is used as one of the basic representations of the human body in video, refer to Fig.1 for example. This volume contains two image axes X and Y, and a temporal axis T, therefore it can encode both the appearance and motion information of the human body. Unlike the previous works [24,23] which extract the appearance and motion information as two separate channels, in this work, the moving human body is considered as a 3D entity in the spatial-temporal domain. This 3D entity comes from the motion of the contours/edges. When a contour/edge in the image plane moves along the temporal axis, its trajectory will extend a surface in the spatial-temporal domain. Take Fig.2 for example, the line L in frame I_0 translates to the line L' in frame I_{T-1} through a uniform linear motion. Its trajectory from frame



(a) A moving line **L** in the image sequence generates a plane **S** in the space-time volume

(b) The plane **S** corresponds to a point **P** in the Hough space that parameterized by (ϕ, θ, ρ)

Fig. 2. The 3D planes that generated by the human motion and its mapping in the Hough space

I_0 to I_{T-1} can expand a plane S. When the contour/edge is not linear or the motion is not uniform, the plane will change to a surface, called by us the spatial-temporal surface. Therefore, the moving human body can be considered as the combination of many spatial-temporal surfaces.

There are two challenges for this surface based human representation: firstly, since the contours/edges in the real-world images are usually not well defined geometrical structures and the motions of the human body are usually complex, the spatial-temporal surfaces may not be in any well defined geometrical forms and can not be explained analytically; secondly, due to the imperfections in either the image data or the edge detector, the contours/edges in the images may not be smooth and continuous. Therefore the smoothness and continuity of the spatial-temporal surfaces can not be guaranteed.

For the first challenge, a possible solution is to use the combination of smaller 3D facets to approximate the surfaces with arbitrary structure. In this way, the 3D planes (facets) are further introduced as the primitives to represent the spatial-temporal surface, and the moving human body can be parsed as a combination of these planes. Regarding the second challenge, we extend the definition of the plane to make it to tolerate the discontinuity using spatial and angular uncertainty. This relaxed definition of the plane is referred to as *generalized plane*, in which the uncertainty of the rotation and translation are defined explicitly. More details will be presented in the following sections.

3 The Definition of the Generalized Plane

In the 3D spatial-temporal domain, a plane is represented by its explicit equation as $t = ax + by + c$. Here, we can use a 3D Hough space corresponding to the parameters a, b and c. However, this formulation suffers from the following problem: as the planar direction becomes vertical, the values of some parameters will become too big and even infinite. This means some planes are not well defined in this $a - b - c$ Hough space.

To avoid the above problem, we parameterize the plane by its normal direction $\boldsymbol{n} = (n_x, n_y, n_t)$ and its perpendicular distance ρ from the origin instead, as in Fig.2(a). This is also called Hesse normal form of the plane, and can be represented as follows:

$$\rho = \boldsymbol{p} \cdot \boldsymbol{n} \tag{1}$$

where $\boldsymbol{p} = (x, y, t)$ is the coordinates of the points on the plane. As there is a constraint on the magnitude of the normal of the plane, i.e. $||\boldsymbol{n}|| = 1$, there are only two degrees of freedom for $\boldsymbol{n} = (n_x, n_y, n_t)$. Therefore, the normal direction $\boldsymbol{n}$ can be represented by the spherical coordinates of a unit sphere (ϕ, θ) as:

$$\boldsymbol{n} = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi) \tag{2}$$

where the inclination $\phi \in [0, \pi]$ is the angle between the zenith direction and $\boldsymbol{n}$; the azimuth $\theta \in [0, 2\pi)$ is the angle between the reference direction on the chosen plane and the projection of $\boldsymbol{n}$ on the plane, as shown in Fig.3(b).

Therefore, by replacing the Equ.2 into the Equ.1, we can get the representation of the plane in the spherical coordinates as:

$$\rho = x\sin\phi\cos\theta + y\sin\phi sin\theta + t\cos\phi \tag{3}$$

In this definition, there are three parameters ϕ, θ and ρ, and the Hough space can be defined accordingly. We refer to this Hough space as the $\phi - \theta - \rho$ Hough space and all the planes can be well defined in this space. Any plane S in the STV space can map to a point P in this Hough space, as shown in Fig.2(b). Any point (x, y, t) on this plane satisfies the definition:

$$\big\{(x, y, t)|\rho = F(x, y, t; \phi, \theta), (x, y, t) \in \chi^3)\big\} \tag{4}$$

where $F(x, y, t; \phi, \theta)$ is the plane's representation in the spherical coordinates as in Equ.3 and χ^3 denotes the range of the definition of the coordinate (x, y, t).

Theoretically, there is a one-to-one mapping between the planes in the STV space and the points in the Hough space with ϕ, θ and ρ as axes. Taking Fig.2 for example, a plane S in the STV space corresponds to a point $P(\phi_0, \theta_0, \rho_0)$ in the Hough space.

But as mentioned in previous section, for many applications in image processing and computer vision, we can seldom find a plane that strictly meets the geometry definition as in Equ.4 due to the imperfections in either the image data or the edge detector. In addition, due to the translation and rotation uncertainty, a nonideal plane in the STV space evidently does not occupy a single point in the Hough space but a cluster of points instead, as mentioned in [22]. In order to make the definition to accommodate these nonideal planes, we extend the definition of the planes in the Hough space by extending a single point $P(\phi_0, \theta_0, \rho_0)$ into a cuboid region R parameterized by the center position $(\phi_0, \theta_0, \rho_0)$ and the cuboid size $(2\tau_\phi, 2\tau_\theta, 2\tau_\rho)$. This means that all the facets that fall into this cuboid region in the Hough space will still be considered as a *plane*. This *plane* is not a conventional plane that can fulfill the restriction in Equ.4, but it is a

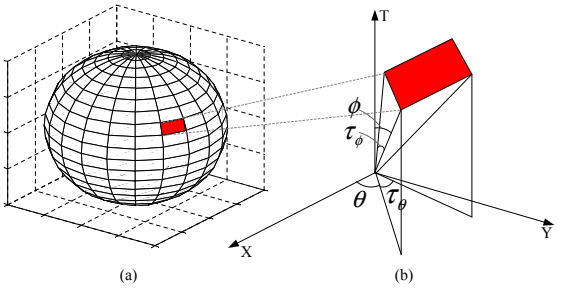


Fig. 3. The 3D orientation partition based on sphere polar coordinates

plane that can tolerate certain degree of rotation and translation uncertainty. This motivates us to generalize the definition of the plane as:

$$\begin{aligned}&\{(x,y,t)|\rho = F(x,y,t;\phi,\theta),(x,y,t)\in\chi^3,\\&|\rho-\rho_0|\le\tau_\rho,|\phi-\phi_0|\le\tau_\phi,|\theta-\theta_0|\le\tau_\theta)\}\end{aligned} \tag{5}$$

We refer to this definition as a *generalized plane*. The geometrical explanation of this definition is that a generalized plane can be a combination of facets that fall into a cuboid region in the Hough space. This endows the generalized plane with robustness to the uncertainty of rotation and translation. Three important properties of the generalized plane are summarized here:

1. It can represent the nonideal planes which can be discontinuous and even with certain level of rotation and translation uncertainty.
2. The robustness to the uncertainty of rotation and translation can be controlled quantitatively by the parameters $(\tau_\phi, \tau_\theta, \tau_\rho)$. More specifically, the robustness to rotation can be controlled quantitatively by (τ_ϕ, τ_θ) and we refer to it as the *rotation uncertainty*; the robustness to translation can be controlled by τ_ρ and we refer to it as the *translation uncertainty*.
3. When we restrict the window size to zero, i.e. $\tau_\phi = 0$, $\tau_\theta = 0$,and $\tau_\rho = 0$, then the *generalized plane* can degenerate into normal plane as defined in Equ.4. Therefore, the *generalized plane* can be considered as a superset of the plane.

The advantage of Equ.5 is that it can incorporate the uncertainty control into the plane's definition explicitly. Therefore, we can produce planes with different description characteristics by varying the uncertainty parameters that are specified by $(\tau_\phi, \tau_\theta, \tau_\rho)$.

4 Orientation-Space Partition in the STV Space

According to the description in section 3, the generalized plane is defined in the Hough space and can be considered as a $2\tau_\phi \times 2\tau_\theta \times 2\tau_\rho$ cuboid region that

centered at $(\phi_0, \theta_0, \rho_0)$. However, the description of this generalized plane is in the STV space. Therefore, we need to back-project the cuboid region from the Hough space into the STV space.

Intuitively, the back-projection of this cuboid region in the STV space is a sandglass-shaped region. We refer to this region in the STV space as a *partition* to distinguish it from the cuboid region in the Hough space. Based on this extension, we can find a one-to-one mapping between a cuboid region in the Hough space and a partition in the STV space. We achieve this goal by orientation partition and space partition.

4.1 Orientation Partition

Orientation partition is the back-projection of the angular uncertainty (τ_ϕ, τ_θ) from the Hough space to the STV space. As we have mentioned in previous section, the normal direction of the plane is determined by the parameters ϕ and θ, and they have very specific meanings in the spherical coordinates: the inclination ϕ is the angle between the zenith direction and the normal direction; the azimuth θ is the angle between the reference direction on the chosen plane and the projection of the normal direction on the plane. The space expanded by ϕ and θ can be represented on a unit sphere, as shown in Fig.3. There is a one to one mapping between the points on this sphere and the unit directional vectors. Therefore, the partition on this unit sphere corresponds to a partition on the orientation space. We apply the 2-dimensional quantization on the unit sphere by step size τ_ϕ and τ_θ, as shown in Fig.3(a). Thus, the unit sphere is divided into a group of disjoint patches, and the directional vectors that map to the same path are quantized to the same direction. By this means, the uncertainty parameters τ_ϕ and τ_θ can be mapped from the Hough space to the STV space.

More specifically, given a point (x, y, t) on the spatial-temporal volume V, the first-order derivatives (using filter $[1, 0, -1]$) of the intensity along the three directions are represented as (V_x, V_y, V_t). Then the normal direction of this point can be calculated as:

$$\begin{cases} n_x = V_x/s \\ n_y = V_y/s \\ n_t = V_t/s \end{cases} \tag{6}$$

where $s = \sqrt{V_x^2 + V_y^2 + V_t^2}$ is the strength of the gradient. The orientation parameters ϕ and θ can be calculated as:

$$\begin{cases} \theta = \arctan n_y/n_x \\ \phi = \arctan \sqrt{(n_x^2 + n_y^2)/n_t^2} \end{cases} \tag{7}$$

By Equ.3, we can calculate the distance parameter ρ also. Therefore, any point on the STV can be represented by a septet $[x, y, t, s, \phi, \theta, \rho]$. Then we quantize the angles ϕ and θ by step size τ_ϕ and τ_θ respectively, according to the rotation

uncertainty defined in Equ.5. Thus far, the original STV has been divided into $m \times n$ disjoint directional channels:

$$\left\{[x, y, t, s, \rho]_{\phi_1-\theta_1}, \ldots, [x, y, t, s, \rho]_{\phi_i-\theta_j}, \ldots, \ [x, y, t, s, \rho]_{\phi_m-\theta_n}\right\} \tag{8}$$

where:
m, n — number of index by quantization, $m = \lceil \pi/\tau_\phi \rceil$, $n = \lceil \pi/\tau_\theta \rceil$;
ϕ_i, θ_j — the quantized inclination and azimuth angles, $\phi_i = i * \tau_\phi$, $\theta_j = j * \tau_\theta$;
$\phi_i - \theta_j$ — the symbol that represents the principal orientation of each channel;
For each channel $[x, y, t, s, \rho]_{\phi_i-\theta_j}$, only the voxels whose normal angle can be quantized as its principal orientation $\phi_i - \theta_j$ are preserved and all the other voxels are set to zero. We refer to this operation as the orientation partition, as shown in Fig.4(a)-(c).

For simplicity, we will use $V_{\phi_i-\theta_j}$ to denote the channel $[x, y, t, s, \rho]_{\phi_i-\theta_j}$. Therefore, the results of orientation partition can be represented as:

$$\left\{V_{\phi_i-\theta_j} | i = 1, \ldots, m; j = 1, \ldots, n\right\} \tag{9}$$

where $\bigcup_{i=1,j=1}^{i=m,j=n} V_{\phi_i-\theta_j} = V$ and $V_{\phi_i-\theta_j} \bigcap V_{\phi_p-\theta_q} = \emptyset, i \neq q, j \neq q$.

4.2 Space Partition

Space partition is used to back-project the translation uncertainty τ_ρ into the STV space. For each channel $[x, y, t, s, \rho]_{\phi_i-\theta_j}$, it can be further partitioned by a group of parallel planes, and we refer to these planes as the partition planes. Moreover, the normal directions of all the partition planes are equal to the principal direction $\phi_i - \theta_j$ of the current channel and the distances between the adjacent partition planes are equal to the translation uncertainty parameter τ_ρ. The region between the two adjacent partition planes can be considered as a generalized plane and all the voxels located within this region belong to the same generalized plane. By this means, we can explicitly control the plane's robustness to the translation uncertainty.

The space partitions can be now represented as $[x, y, t, s]^{\rho_k}_{\phi_i-\theta_j}$. We denote it by $P^{\rho_k}_{\phi_i-\theta_j}$ for simplicity. These partitions fulfill the following definition:

$$\left\{P^{\rho_k}_{\phi_i-\theta_j} | k = 1, \ldots, o; \bigcup_{k=1}^{o} P^{\rho_k}_{\phi_i-\theta_j} = V_{\phi_i-\theta_j}; \ P^{\rho_m}_{\phi_i-\theta_j} \bigcap P^{\rho_n}_{\phi_i-\theta_j} = \emptyset, m \neq n\right\} \tag{10}$$

where o is the number of spatial partition, as shown in Fig.4(d).

Moreover, the strength of the gradient within each partition can be represented as:

$$g^{\rho_k}_{\phi_i-\theta_j} = q(P^{\rho_k}_{\phi_i-\theta_j}) \tag{11}$$

where $q(\cdot)$ is the function that calculates the summation of gradient strength within a partition.

By the orientation and space partition, we can associate a cuboid region in the Hough space with a partition in the STV space. The representation property of the generalized plane can be controlled by $(\tau_\phi, \tau_\theta, \tau_\rho)$ during the partition procedure. The overall orientation-space partition procedure can be seen in Fig.4. The statistical description of the generalized plane can be calculated easily within each partition, which will be detailed in the following section.

5 Computation of STGGP Descriptor

Thus far, we have mapped the generalized plane from the Hough space to the STV space by orientation-space partition. In this section, we will present how to calculate the descriptors for the generalized plane in the STV space.

The descriptor of the generalized plane is 9-dimensional heterogeneous vector. It can encode the gradient strength, position and shape information of the plane. For any channel $V_{\phi_i-\theta_j}$, its descriptor can be represented as:
$(i_{max}, g_{max}, \sigma, m_x, m_y, m_t, v_{norm}, v_{tangX}, v_{tangY})_{\phi_i-\theta_j}$.

Given a feature that is specified by a cuboid $C(x_0, y_0, t_0, w, h, l)$ and the uncertainty parameters $(\tau_\phi, \tau_\theta, \tau_\rho)$, we firstly perform the orientation-space partition within C as mentioned above. Then for each channel $V_{\phi_i-\theta_j}$, we can get the space partitions $P^{\rho_k}_{\phi_i-\theta_j}$ as Equ.10. The gradient strength $g^{\rho_k}_{\phi_i-\theta_j}$ of each partition can be calculated as in Equ.11. In the following section, we will drop the orientation subscript $\phi_i - \theta_j$ for simplicity. The items of the descriptor can be calculated as follows:

- i_{max}: is the normalized index value of the space partition with the maximum gradient strength. It can be calculated as $i_{max} = \frac{i'_{max}}{o}$, where $i'_{max} = \arg\max_k (g^{\rho_k})$ is the index of the space partition with the maximum gradient strength and o is the number of space portions.
- g_{max}: is the normalized maximum gradient strength, where $g_{max} = \frac{g'_{max}}{\sum_{k=1}^{o} g^{\rho_k}}$, $g'_{max} = \max(g^{\rho_k})$.
- σ: is the standard deviation of the gradient strength, and can be calculated as $\sigma = \sqrt{\frac{1}{o}\sum_{k=1}^{o}(g^{\rho_k} - \bar{g})^2}$, where $\bar{g} = \frac{1}{o}\sum_{k=1}^{o} g^{\rho_k}$.
- (m_x, m_y, m_t): is the normalized mean position of all the non-zero points in partition $P^{\rho_{i'_{max}}}$, for example, m_x is calculated as follows:

$$m_x = \frac{1}{z}\sum_{i=0}^{z}\frac{(x_i - x_0)}{w} \tag{12}$$

  where:
  z — the number of non-zero points in the partition $P^{\rho_{i'_{max}}}$;
  x_0 — the center point of the feature cuboid C;
  w — the size of the feature cuboid C;
- $(v_{norm}, v_{tangX}, v_{tangY})$: denote the standard deviation of the non-zero points in partition $P^{\rho_{i'_{max}}}$, for example, v_{norm} is be calculated as follows:

$$v_{norm} = \sqrt{\frac{1}{z}\sum_{i=1}^{z}(r_{i,norm} - m_{norm})^2} \tag{13}$$

  where $r_{i,norm}$ and m_{norm} are the new position of the points and their means in the rotated coordinates;

The reason for coordinate rotation is to align the normal direction of the generalized plane with the axis of the new coordinate frame. In this new coordinate

frame, the distribution of the non-zero points can be easily described. And the rotation matrix A of the current channel $\phi_i - \theta_j$ is defined as:

$$\begin{aligned} A &= A_{\phi_i} A_{\theta_j} \\ &= \begin{bmatrix} \cos\phi_i \cos\theta_j & -\cos\phi_i \sin\theta_j & -\sin\phi_i \\ \sin\theta_j & \cos\theta_j & 0 \\ \sin\phi_i \cos\theta_j & -\sin\phi_i \sin\theta_j & \cos\phi_i \end{bmatrix} \end{aligned} \tag{14}$$

The new coordinate frame is referred to as $norm - tangX - tangY$. In this new coordinate frame, the generalized plane is parallel to the $tangX - tangY$ plane and its norm is aligned with the $norm$ axis. In this new coordinate, the shape property of the generalized planed can be easily characterized. For example, v_{norm} is the standard deviation along the normal direction, and it can be considered as the "thickness" of the generalized plane; and (v_{tangX}, v_{tangY}) can be used to describe the "shape" of the plane.

Thus far, for each channel $V_{\phi_i - \theta_j}$, we have obtained a 9-dimensional feature vector, i.e., $(i_{max}, g_{max}, \sigma, m_x, m_y, m_t, v_{norm}, v_{tangX}, v_{tangY})_{\phi_i - \theta_j}$. Then, by concatenating the feature vectors of all the channels/orientations, we can get the final STGGP descriptor.

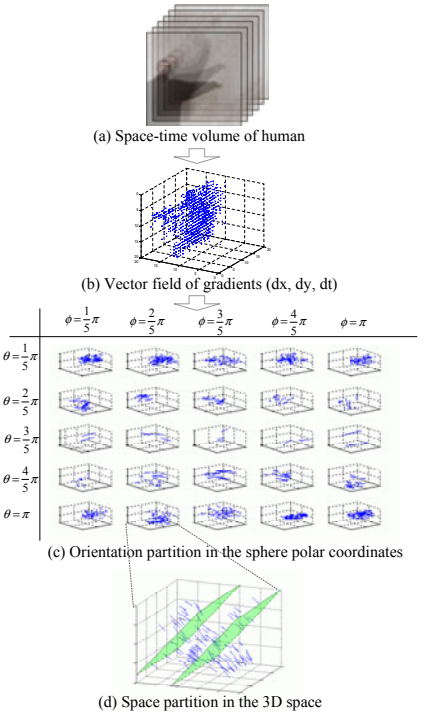


Fig. 4. The overview of the orientation-space partition on the spatial-temporal volume

6 Experiments

In this section, the proposed STGGP is evaluated on the public dataset. Firstly, since STGGP is a heterogeneous vector, we evaluate the contributions of the different components; secondly, we investigate how the temporal length affects the the performance of the method; thirdly, we evaluate the proposed method against the state of the art methods. In our experiments, the linear SVMs are used as the classifiers with parameter $C = 0.1$. For easy comparison, we plot the "recall" vs. "false positive per image" curve.

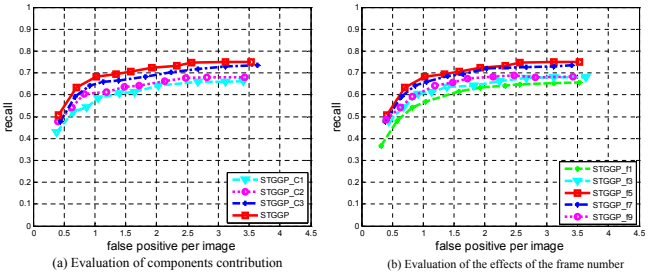
We use the ETHZ as the benchmarking dataset[29]. The size of normalized STV is $96 \times 64 \times 5$ and the size of SGTTP feature is $16 \times 16 \times 5$. With 2 translation uncertainty settings, $\tau \in \{4, 8\}$, we use 78 STGGP features for representing a STV. The orientation uncertainty parameters are set as: $\tau_\phi = \pi/5$ and $\tau_\theta = \pi/5$, therefore, there are 25 orientation channels. The overall dimension of a STV descriptor is 17550.

Since STGGP is a heterogeneous feature vector, it contains the gradients' strength, position and shape information. Therefore, in the first experiment, it is worth to evaluate the contributions of these different components. We reorganized the components of the GGP descriptors as follows:

- $STGGP\_C1$: (g_{max}) only contains the maximum gradients' strength of the partitions, and its description ability is close to HOG.
- $STGGP\_C2$: $(g_{max}, i_{max}, \sigma)$ adds partition index and the standard deviation of the gradient strength to represent the strength distribution information within the feature region.
- $STGGP\_C3$: $(g_{max}, i_{max}, \sigma, m_x, m_y)$ adds the mean positions of all the non-zero pixels to represent the poison information.
- $STGGP$: $(g_{max}, i_{max}, \sigma, m_x, m_y, v_{norm}, v_{tang})$ the STGGP descriptor that adds the standard deviations of positions to represent the shape of non-zero pixels in the partition.

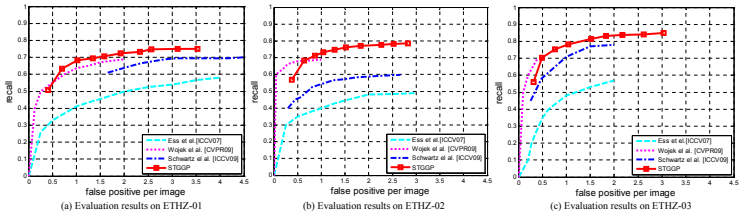
The evaluation results can be seen in Fig.5(a). From these results, we can make a few observations: firstly, the performance can be improved monotonically as long as the new components are heterogeneous to the previous ones; secondly, the position information is critical of the performance, and the most prominent improvements can be observed after the position information have been added.

In the second experiment, we evaluate how the number of frames can affect the performance of STGGP. Therefore, we re-generate the sample sets with different frame numbers and train the detectors from these sets. Here we use $STGGP\_fn$ to represent the detectors that trained with different frame numbers and n is the number of frame. For $STGGP\_f1$, we just use the original GGP detector that use only the static features. The results of these detectors can be seen in Fig.5(b). When we add frame number from 1 to 5, a monotonously improvement can be observed; but when we add more frames, the performance actually dropped. A possible explanation of is that for a longer temporal duration, the ego-motion of the camera will substantially affect the spatial-temporal shape of the human.



(a) Evaluation of components contribution (b) Evaluation of the effects of the frame number

Fig. 5. Parameter evaluation on ETH-01 set



(a) Evaluation results on ETHZ-01 (b) Evaluation results on ETHZ-02 (c) Evaluation results on ETHZ-03

Fig. 6. Evaluation results on the ETHZ human dataset

In the third experiment, the STGGP detector is evaluated against the state of the art methods [29,25,3]. On ETH-01 set, the STGGP yields comparable results to the best results in [25], as shown in Fig.6(a); On ETH-02 and ETH-03 sets, the STGGP outperforms the other methods, as shown in Fig.6(b)(c). Another observation is that the features combining both appearance and motion outperform the appearance only based detector by a big margin. Some samples of the detection results can be found in Fig.7.

Fig. 7. Sample detection results on ETHZ dataset

7 Conclusion

In this paper we have developed a spatial-temporal granularity-tunable gradients partition (STGGP) descriptor to represent the human's motion pattern

in the spatial-temporal domain. Firstly, the *generalized plane* is defined in the Hough space. By incorporating the rotation and translation uncertainties in the definition of the plane, it can describe the object with a family of descriptors with different representation ability, from the detailed geometrical representation to the statistical description. Then, by orientation-space partition, the generalized plane can be back-projected from the Hough space to the spatial-temporal space. Finally, we form the heterogeneous descriptor in the generalized plane. The heterogeneous descriptor contains gradient's strength and spatial distribution information, which further improve its representation ability. The STGGP descriptor is tested for human detection in image sequences and promising results have been achieved.

Acknowledgement

This paper is partially supported by Natural Science Foundation of China under contracts No.60772071, No.60832004, No.60872124, and No. U0835005; National Basic Research Program of China (973 Program) under contract 2009CB320902. The financial support from the Infotech Oulu is also gratefully acknowledged.

References

1. Han, F., Shan, Y., Sawhney, H.S., Kumar, R.: Discovering class specific composite features through discriminative sampling with swendsen-wang cut. In: CVPR (2008)
2. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV, pp. 32–39 (2009)
3. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: ICCV (2009)
4. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: CVPR (2009)
5. Dollar, P., Babenko, B., Belongie, S., Perona, P., Zhuowen, T.: Multiple component learning for object detection. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 211–224. Springer, Heidelberg (2008)
6. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR (2009)
7. Ott, P., Everingham, M.: Implicit color segmentation features for pedestrian and object detection. In: ICCV, pp. 724–730 (2009)
8. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR, pp. 511–518 (2001)
9. Papageorgiou, C., Poggio, T.: A trainable system for object detection. IJCV 38, 15–33 (2000)
10. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. TPAMI 23, 349–361 (2001)
11. Gavrila, D.M.: Pedestrian detection from a moving vehicle. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 37–49. Springer, Heidelberg (2000)
12. Lin, Z., Davis, L.S., Doermann, D., DeMenthon, D.: Hierarchical part-template matching for human detection and segmentation. In: ICCV (2007)

13. Lin, Z., Davis, L.S.: A pose-invariant descriptor for human detection and segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 423–436. Springer, Heidelberg (2008)
14. Ferrari, V., Tuytelaars, T., Gool, L.V.: Object detection by contour segment networks. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 14–28. Springer, Heidelberg (2006)
15. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: ICCV (2005)
16. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV, pp. 1150–1157 (1999)
17. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 69–82. Springer, Heidelberg (2004)
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
19. Zhu, Q., Avidan, S., Yeh, M.C., Cheng, K.T.: Fast human detection using a cascade of histograms of oriented gradients. In: CVPR, pp. 1491–1498 (2006)
20. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
21. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on riemannian manifolds. In: CVPR (2007)
22. Liu, Y., Shan, S., Zhang, W., Gao, W., Chen, X.: Granularity-tunable gradients partition (ggp) descriptors for human detection. In: CVPR, pp. 1255–1262 (2009)
23. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
24. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: ICCV, pp. 734–741 (2003)
25. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: CVPR (2009)
26. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: CVPR, vol. 2, pp. 123–130 (2001)
27. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: CVPR, vol. 2, pp. 1395–1402 (2005)
28. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: ICCV, vol. 1, p. 166–173 (2005)
29. Ess, A., Leibe, B., Gool, L.V.: Depth and appearance for mobile scene analysis. In: ICCV, pp. 14–21 (2007)