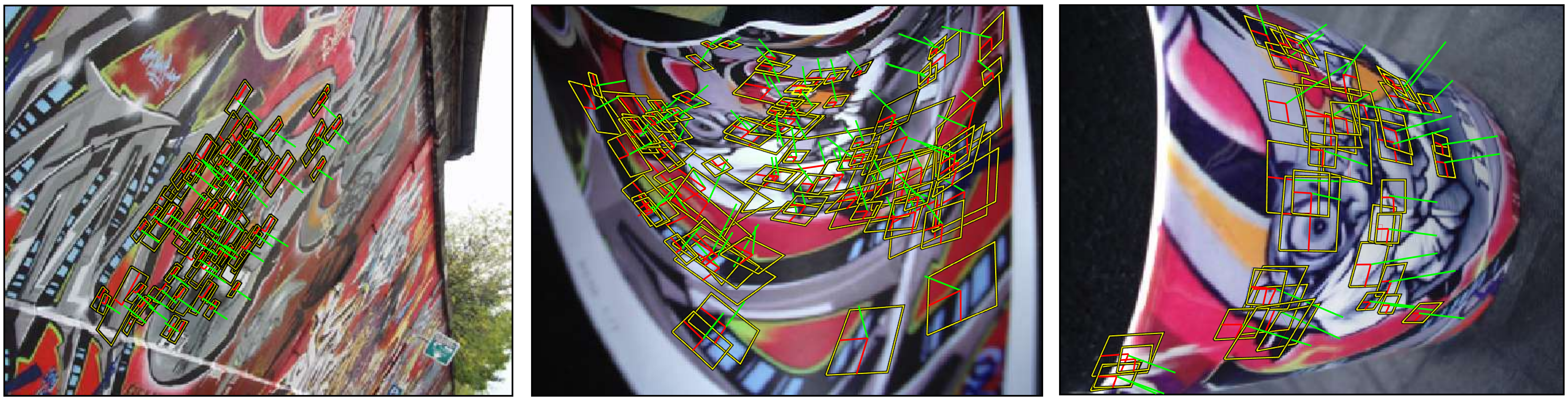
LOCAL 3D SURFACE POSE ESTIMATION BY NUISANCE RESIDUAL LEARNING

Del Bimbo A., Franco F. and Pernici F.

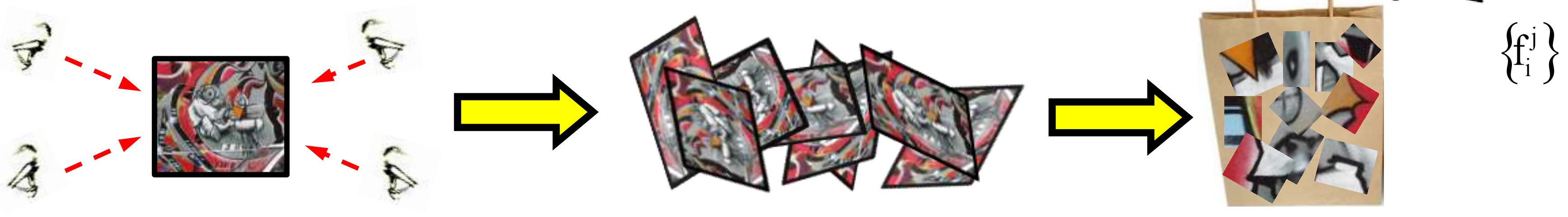
Media Integration and Communication Center (MICC), University of Florence, Italy



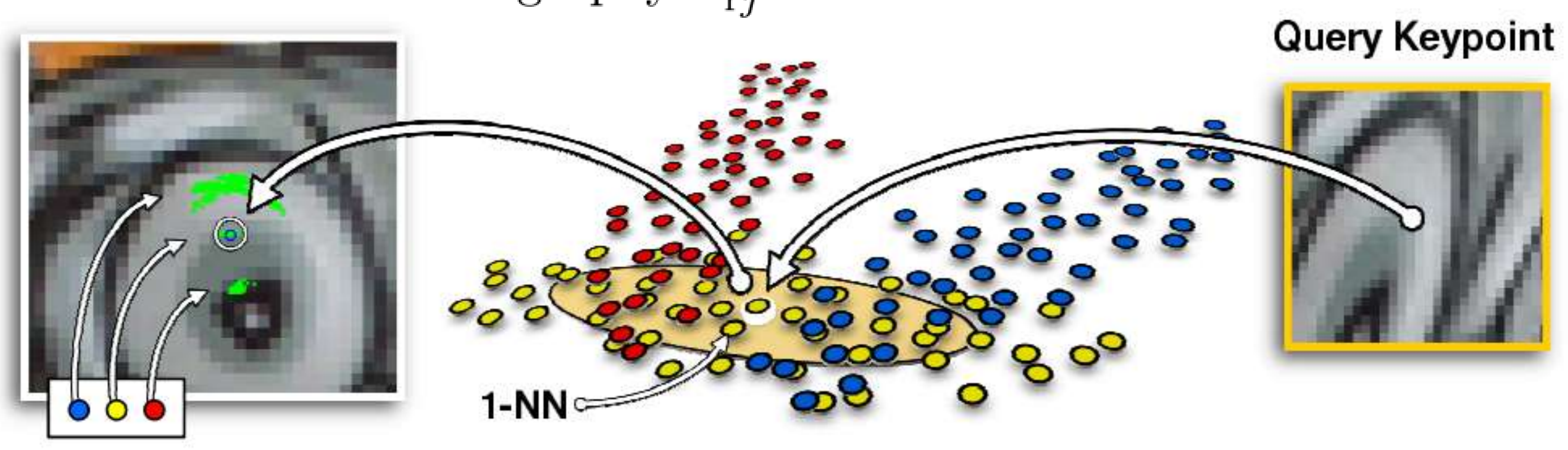
Abstract: We present a method of estimating the pose of an imaged scene surface element provided that it can be locally approximated by its tangent plane. The approach **simultaneously** learn the "nuisance residual" structure present in the detection and description steps of the SIFT algorithm allowing local perspective properties to be recovered through a homography. The estimated local poses can be applied to non rigid surfaces, with an accuracy representative of state-of-the-art for this challenging task.

Given a test image $\mathbf{I}_\mathrm{t}$ of one object taken from an arbitrary unknown point of view and a reference view $\mathbf{I}_\mathrm{r}$ of the object from which are taken multiple views with their homographies, we address the problem of finding the point of view from which $\mathbf{I}_\mathrm{t}$ is taken. In particular our goal is to estimate the homography $\mathbf{h}_\mathrm{rt}$ between $\mathbf{I}_\mathrm{r}$ and $\mathbf{I}_\mathrm{t}$, regardless of scale and perspective distortion that possibly affect the test image. To this end we assume that the surface of the 3D object has smooth curvature, so that the region around a keypoint can be considered locally planar.

Training set generation. Given a reference image $\mathbf{I}_\mathrm{r}$ of a planar surface, we take a full set of images taken from different viewpoints $\mathbf{I}_j$ and extract SIFT keypoints in each view. Each keypoint $\mathbf{k}_i^j$ in $\mathbf{I}_j$ is associated with a representation feature that accounts for SIFT description, location, main orientation, scale at which it is detected, and the homographic transformation $\mathbf{h}_{\mathrm{r}j}^i$ between a reference view $\mathbf{I}_\mathrm{r}$ and $\mathbf{I}_j$: $\mathbf{f}_i^j = \{(x_i^j, y_i^j), s_i^j, o_i^j, \mathbf{d}_i^j, \mathbf{h}_{\mathrm{r}j}^i\}$.



Matching and geometry consistency check. The set $\mathbf{C}^\mathrm{t} = \{\mathbf{f}_i^j \in \mathcal{F} : i \in N_k(\mathbf{d}^\mathrm{t})\}$ of corresponding features of keypoint $\mathbf{k}^\mathrm{t}$ is retrieved as the k-nearest neighbours N_k in the space of SIFT appearance descriptors. Outliers are removed by back-projecting $\mathbf{f}_i^j \in \mathbf{C}^\mathrm{t}$ in the coordinate system of $\mathbf{I}_\mathrm{r}$ using the inverse of its own homography $\mathbf{h}_{\mathrm{r}j}^{i\,-1}$.



Exploiting detection nuisance. In order to obtain geometrically meaningful homographies from the dataset with respect to a test keypoint $\mathbf{k}^\mathrm{t} = \{(x^\mathrm{t}, y^\mathrm{t}), s^\mathrm{t}, o^\mathrm{t}, \mathbf{d}^\mathrm{t}\}$ and to **simultaneously** take into account the "residual" invariance of the SIFT detector and descriptor algorithm, the support region of retrieved features in $\mathbf{C}^\mathrm{t}$ are aligned. We calculate the shifts of: position $(u_i^j, v_i^j) = (x^\mathrm{t} - \mu_x, y^\mathrm{t} - \mu_y)$, and scale $(\sigma_i^j = \frac{s^\mathrm{t}}{s_i^j})$, and orientation $(\theta_i^j = o^\mathrm{t} - o_i^j)$ between features $\mathbf{f}_i^j \in \mathbf{C}^\mathrm{t}$ and $\mathbf{k}^\mathrm{t}$, being (μ_x, μ_y) the centroid of keypoints in $\mathbf{C}^\mathrm{t}$. The similarity transformation $\mathbf{s}_{j\mathrm{t}}^i = \begin{bmatrix} \sigma_i^j \cos\theta_i^j & -\sin\theta_i^j & u_i^j \\ \sin\theta_i^j & \sigma_i^j \cos\theta_i^j & v_i^j \\ 0 & 0 & 1 \end{bmatrix}$ accounts for scaling, translation and rotation and can be used to align the neighbourhoods of features $\{\mathbf{f}_i^j\}$ to the region of input keypoint $\mathbf{k}^t$, through the following homographic transformation: $\mathbf{h}_{\mathrm{r\,t}}^i = \mathbf{h}_{\mathrm{r}j}^i \cdot \mathbf{s}_{j\mathrm{t}}^i$

Learning with kernel regression. Let $\mathbf{D} \in \mathbb{R}^{128}$ denotes a real valued random vector descriptor, and $\mathbf{H} \in \mathbb{R}^8$ a real valued random output homographic transformation, with joint distribution $p_{\mathbf{D},\mathbf{H}}(\mathbf{d}, \mathbf{h})$. Given a training set consisting of N input-output pairs: $\{(\mathbf{d}_i^j, \mathbf{h}_\mathrm{rt}^i)_1, ..., (\mathbf{d}_i^j, \mathbf{h}_\mathrm{rt}^i)_\mathrm{N}\}$ for which the probability of taking a specific value obeys $p_{\mathbf{D},\mathbf{H}}(\mathbf{d}, \mathbf{h})$, we wish to learn the function $\Phi : \mathbf{D} \mapsto \mathbf{H}$ defined as $\mathbf{h} = \Phi(\mathbf{d}) = [\Phi_1(\mathbf{d}), ..., \Phi_8(\mathbf{d})]$ that maps from the space of input vector descriptors to the space of homographies. According to this, we can define a local regression estimate of $\Phi(\mathbf{d}^\mathrm{t})$ as $\Phi_{\hat{\theta}}(\mathbf{d}^\mathrm{t})$, which minimizes the cost: $L(\mathbf{h}_\mathrm{rt}^i, \Phi_\theta(\mathbf{d}^\mathrm{t})) = \sum_{\mathbf{d}_i^j \in \mathbf{C}^\mathrm{t}} K_\Sigma(\mathbf{d}^\mathrm{t}, \mathbf{d}_i^j)[\mathbf{h}_\mathrm{rt}^i - \Phi_\theta(\mathbf{d}_i^j)]^2$ and Φ_θ is some parameterized function. We used the constant function model: $\Phi_\theta(\mathbf{d}) = \theta_0$ in which case the final form of the kernel estimation is the Nadaraya-Watson weighted average: $\mathbf{h}_\mathrm{rt} = \hat{\Phi}(\mathbf{d}^\mathrm{t}) = \hat{\theta}_0 = \frac{\sum_i^\mathrm{N} K_\Sigma(\mathbf{d}^\mathrm{t}, \mathbf{d}_i^j)\mathbf{h}_\mathrm{rt}^i}{\sum_i^\mathrm{N} K_\Sigma(\mathbf{d}^\mathrm{t}, \mathbf{d}_i^j)}$. The kernel K_Σ embeds feature descriptors of a cluster data into a vector space. Local neighborhood metric is specified by a Gaussian kernel function centered on the query descriptor: $K_\Sigma(\mathbf{d}^\mathrm{t}, \mathbf{d}_i^j) = ((2\pi)^n|\Sigma|)^{\frac{1}{2}} \exp^{-\frac{1}{2}(\mathbf{d}_i^j - \mathbf{d}^\mathrm{t})^T \Sigma^{-1} (\mathbf{d}_i^j - \mathbf{d}^\mathrm{t})}$ where $\Sigma = \mathrm{N}^{-1} \sum_{i=1}^\mathrm{N} (\mathbf{d}_i^j - \mathbf{d}^\mathrm{t}) \cdot (\mathbf{d}_i^j - \mathbf{d}^\mathrm{t})^T$.

Experimental Results:

Several experiments were performed in order to assess the effectiveness of the method and compare it against "Leopar" and "Caspar" [1] [2]. In the first set of experiment we compare the average overlap between the quadrangles obtained with our method and those obtained using the ground truth homography (Fig. 1). Similarly to "Leopar" this overlap is very close to 100% for our method. The second set shows the comparison of the mean reprojection error for the quadrangle corners (Fig. 2). The error of the patch corner is less than one pixel in average and outperforms other methods. In the last set we prove the stability of our method with respect to the number of multiview descriptor-homography pairs used in the Kernel based regression (Fig. 3). Error bars show that we obtain a mean reprojection error of less than 1 pixel over almost the range from 30 to 250 descriptors. It also shows that the minimum error is obtained in the range from 140 to 165 descriptor-homography pairs. Our current implementation runs at about 15 frame per second using 400.000 features in the database and extracting about 200 SIFT keypoints in the input images, on a standard notebook with an Intel Centrino Core Duo with 2.4GHz and 3Gb RAM. The average times for the most expensive steps are shown in the table and compares favorably with state-of-the-art methods.

step	*average time*
SIFT Point Extraction	0.025 sec
Geometry Consistency Check	0.021 sec
Kernel Based Regression	0.036 sec

References: **[1]** S. Hinterstoisser, S. Benhimane, N. Navab, P. Fua, and V. Lepetit, "Online learning of patch perspective rectification for efficient object detection," in CVPR, 2008. **[2]** A. Pagani and D. Stricker, "Learning local patch orientation with a cascade of sparse regressors" in BMVC, 2009.

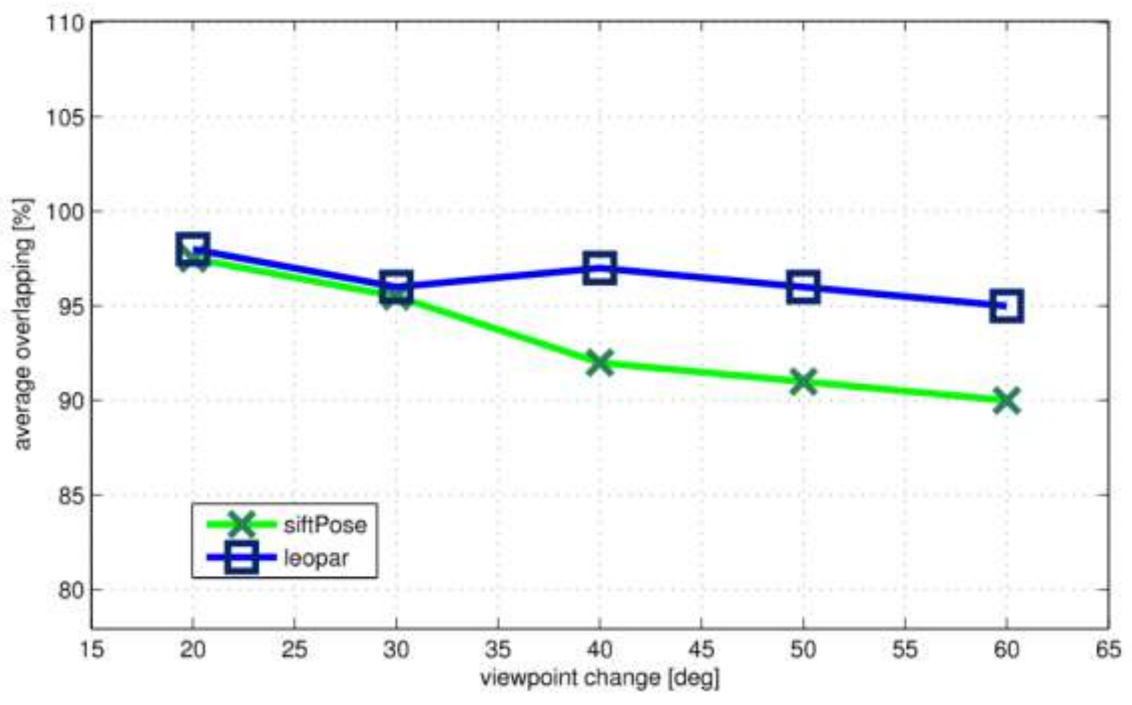


Figure 1

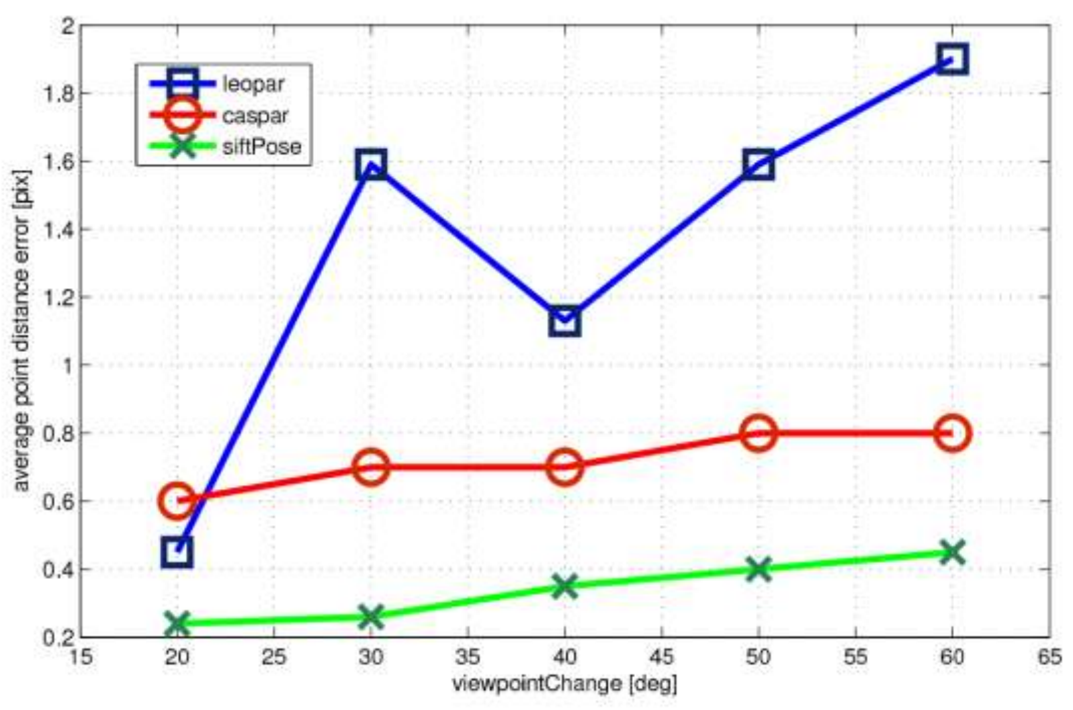


Figure 2

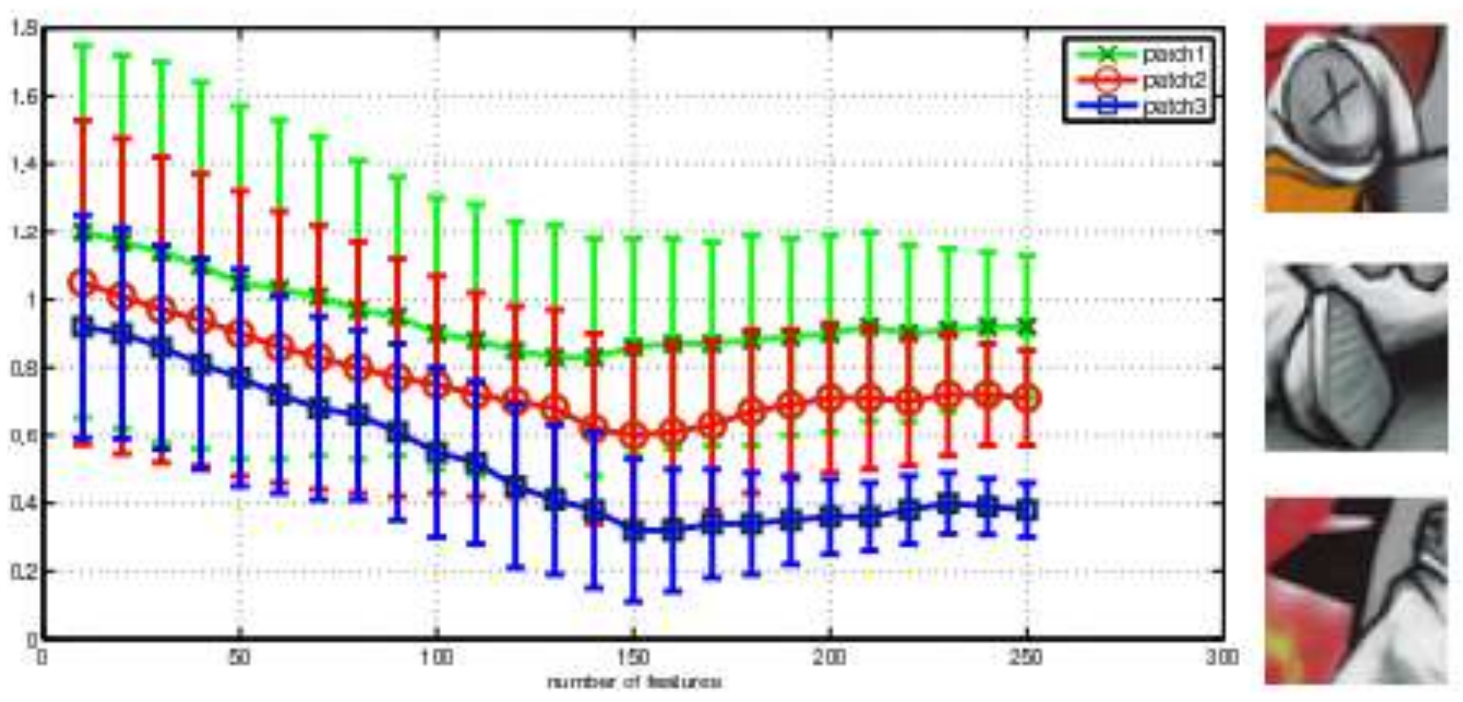


Figure 3