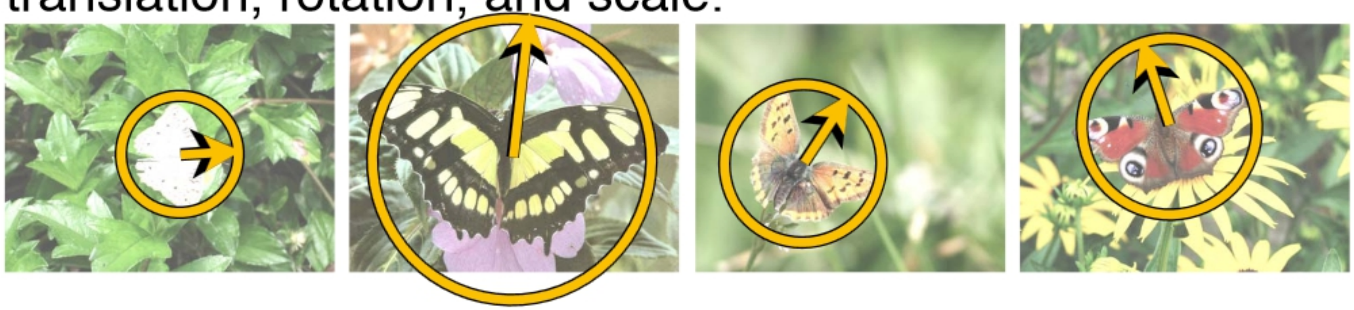
Learning Equivariant Structured Output SVM Regressors

by Vedaldi, Blaschko, Zisserman. ICCV 2011.
Interpreted by Magnus Burenius, KTH.

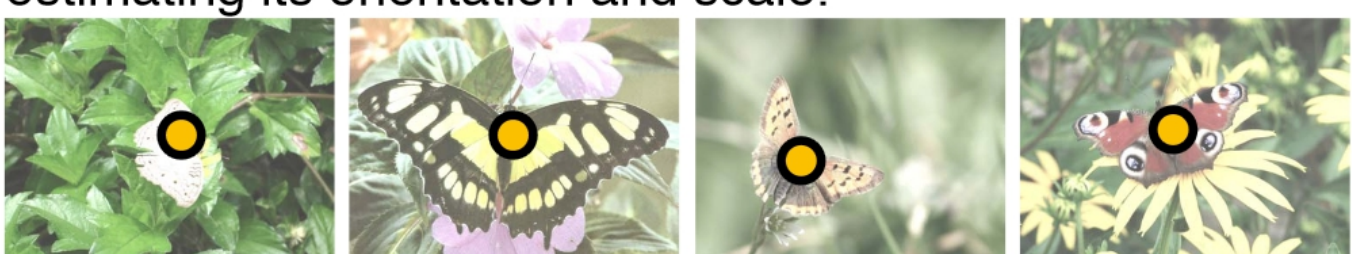
Invariance to Transformations

- **Pose-invariant classification.** Recognize an object category *regardless* of the object translation, rotation, and scale.

- **Pose regression.** Detect an object and estimate its translation, rotation, and scale.

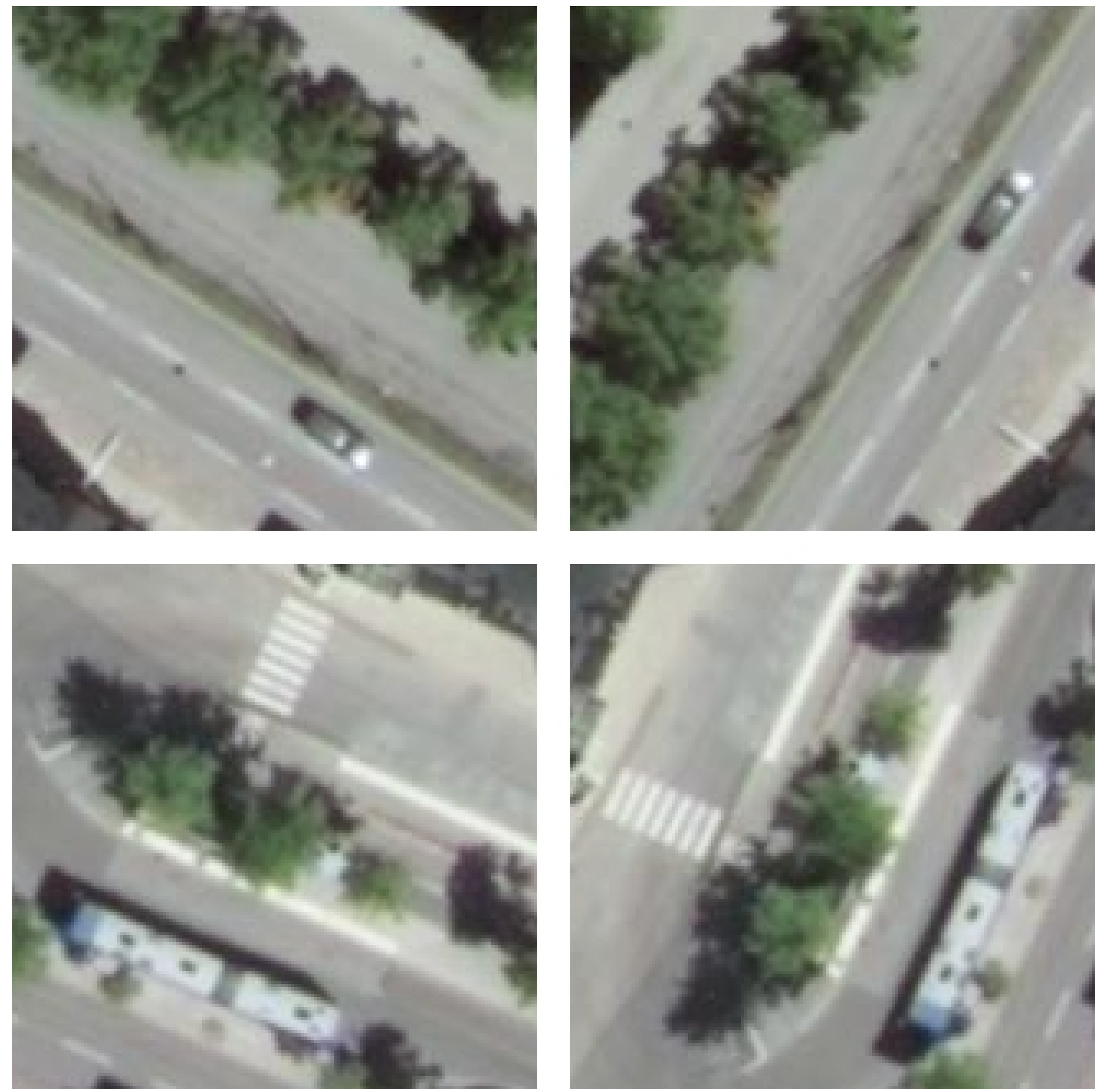


- **Detection.** Find an object location (center), *without* estimating its orientation and scale.



Invariance and Equivariance

- Consider some transformation, like rotation.
- We would like object classification to be invariant to rotation.
- We would like object detection to be equivariant to rotation.



The Problem

- Input space: X
- Output space: Y
- Consider a transformation t acting on the input and output:

$$t=\left(t_X, t_Y\right)\in T$$
$$t_X : X \rightarrow X$$
$$t_Y : Y \rightarrow Y$$

- We want to learn a predictor f(x,w) that is invariant or equivariant to the transformation t:

$$f\left(t_X x ; w\right)=f\left(x ; w\right) \qquad invariance \ \left(t_y=I\right)$$
$$f\left(t_X x ; w\right)=t_Y f\left(x ; w\right) \quad equivariance$$

The Problem

- Most of the time we do not have enough training data, representing all possible transformations.

One Approach

- Can generate more training data by transforming the original data.
- How many samples should be generated? How densely? What samples are relevant?

Another Approach

- Could explicitly model and estimate transformations as latent variables.
- Learning problem becomes non-convex. Inference might be slower.

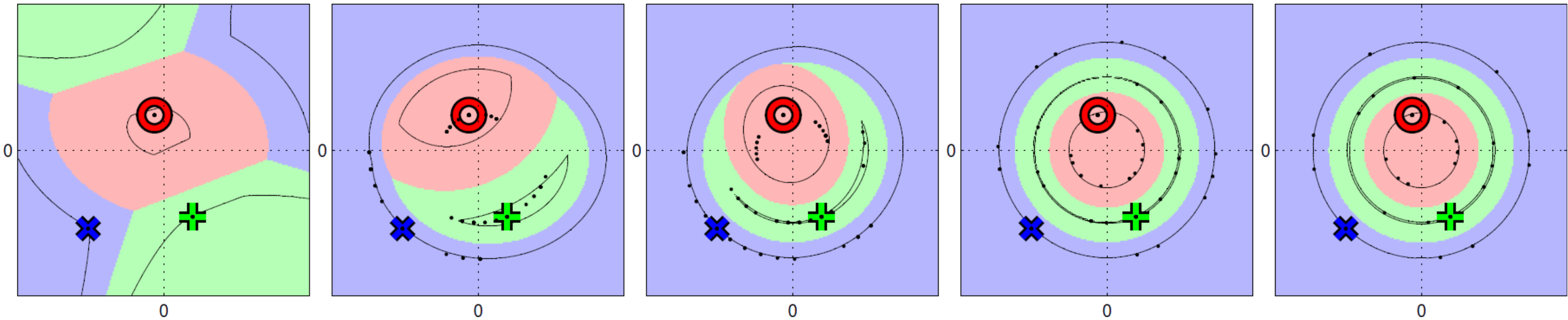
Their Approach

- Generalize Structured SVM to incorporate invariance and equivariance into a convex training procedure.
- Removes the need for ad-hoc sampling strategies. Only generates the virtual samples that are necessary.
- Inference does not require the estimate of latent variables.

Toy Example

Assuming rotation invariance

$$X = R^2$$
$$Y = \{r, g, b\}$$



Gradually enforce invariance to larger rotations →

Standard Structured SVM

- Let X and Y be the input and output and let Ω_X and Ω_Y be their sample spaces. These can be ANY spaces, not just integers or real vector spaces.

- A feature function Ψ is used to map a pair from these complicated spaces to something we can compute with:

$$\Psi : \Omega_X \times \Omega_Y \rightarrow R^d$$

Standard Structured SVM

- A classifier described by a vector ω predicts a class by solving

$$f(x;\omega) = \underset{y}{argmax}\ \ \omega \cdot \Psi(x,y)$$

- This imposes a restriction on Ψ

Standard Structured SVM

During training the STRUCTURE of the output space is taking into account by defining a loss function

$$\Delta : \Omega_Y \times \Omega_Y \rightarrow R$$

which quantifies the loss of predicting y_p when the true output is y. It should fulfill

$$\Delta\left(y, y_p\right) \geqslant 0$$

$$\Delta\left(y, y_p\right) = 0 \quad iff \quad y = y_p$$

Δ should thus reflect the quantity which measures how well the classifier performs.

Standard Structured SVM

- Given a training set $(x_1,y_1)...(x_N,y_N)$ of **”only positives”** and a regularization constant C a classifier ω is trained by solving the convex optimization problem:

$$\min_{\omega} \ \|\omega\|^2 + C \sum_n \max_y \left(\Delta\left(y_n, y\right) + \omega \cdot \Psi\left(x_n, y\right) - \omega \cdot \Psi\left(x_n, y_n\right)\right)$$

Search for difficult classifications

Their Generalization of S-SVM

Standard S-SVM:

$$\min_{\omega} \|\omega\|^2 + C \sum_n \max_y \left(\Delta\left(y_n, y\right) + \omega \cdot \Psi\left(x_n, y\right) - \omega \cdot \Psi\left(x_n, y_n\right)\right)$$

Transformation equivariant generalization:

$$\min_{\omega} \|\omega\|^2 + C \sum_n \max_{y,t} \left(\Delta\left(t, y_n, y\right) + \omega \cdot \Psi\left(t_X x_n, y\right) - \omega \cdot \Psi\left(t_X x_n, t_Y y_n\right)\right)$$

When looking for the difficult classifications
we search over all possible equivariant
variations of input and output.

Training

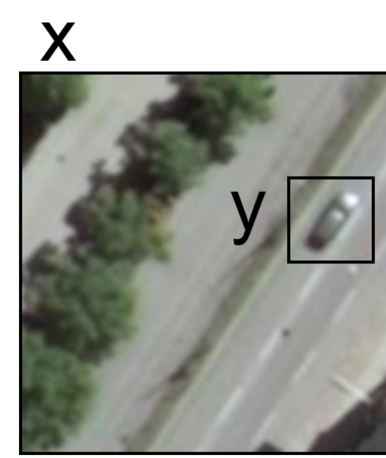
- The problem can be optimized using standard S-SVM solvers.
- These solvers handle the large number of constraints by generating the necessary ones on the fly.
- This corresponds to generating relevant virtual training data.

Advantages

- Principled approach to the generation of relevant virtual training data.
- Training is convex and no more expensive than standard Structured SVM and latent SVM.
- Inference is faster than latent SVM, since the latent variable, corresponding to the transformation, is not estimated.

Experiment 1
Rotation Equivariant Object Detection

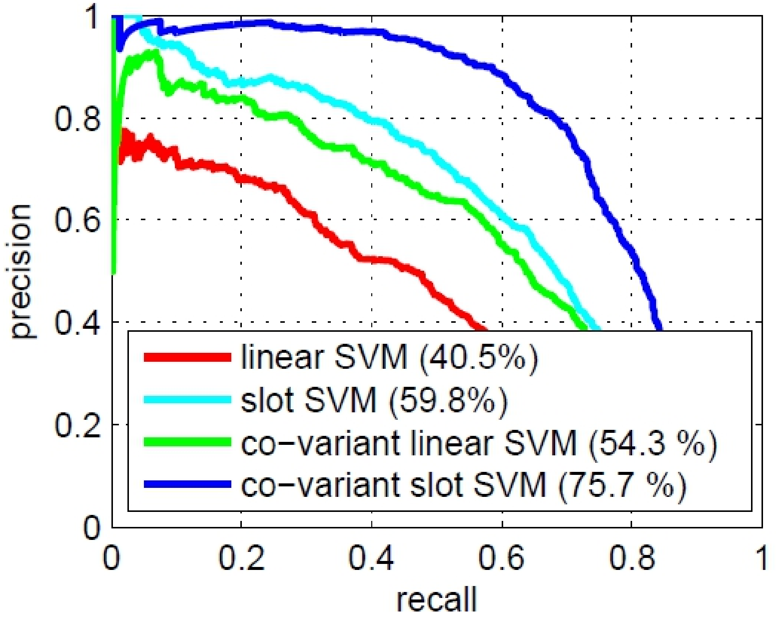
- Let Φ(x,y) be the HOG-descriptor of a block of 7x7 HOG-cells at position y in the image x.
- A linear HOG model is not sufficient to capture arbitrary object rotations.
- They use something they call “slot kernel”.
- The cluster the HOG-space into Q=18 clusters.
- The total feature function is the outer product:

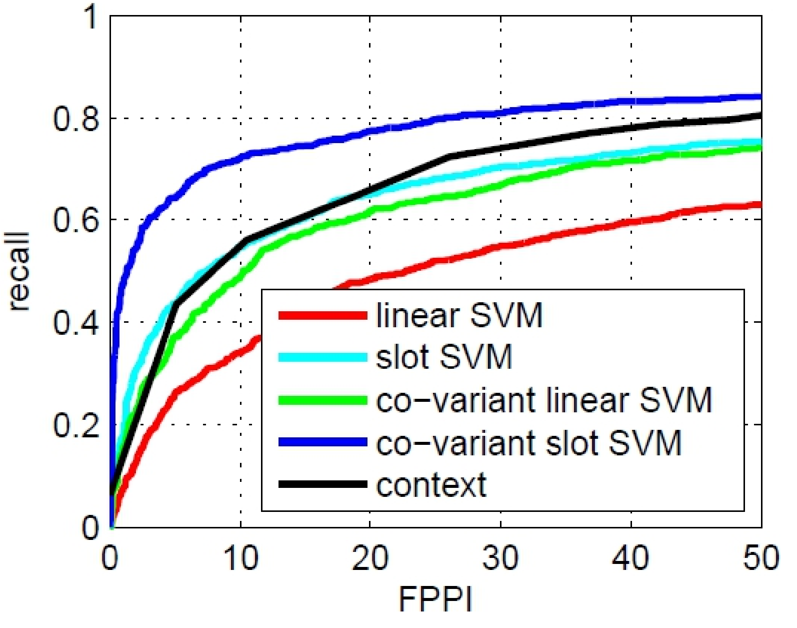


$$\Psi(x,y)=\phi(x,y)\,e^{T}_{q(\phi(x,y))}$$

Experiment 1 - Results

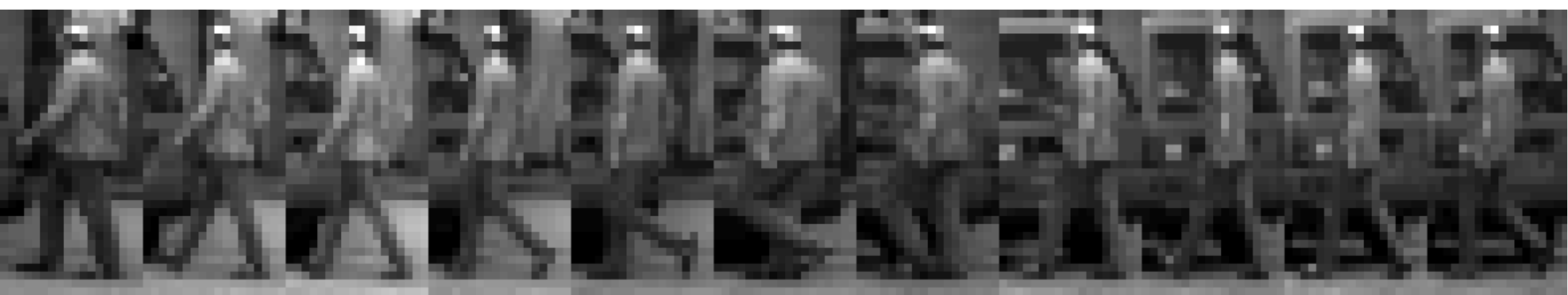
Aerial car detection.
30 images having a total of 1000 cars with different rotations.
Unclear division of training and test data.





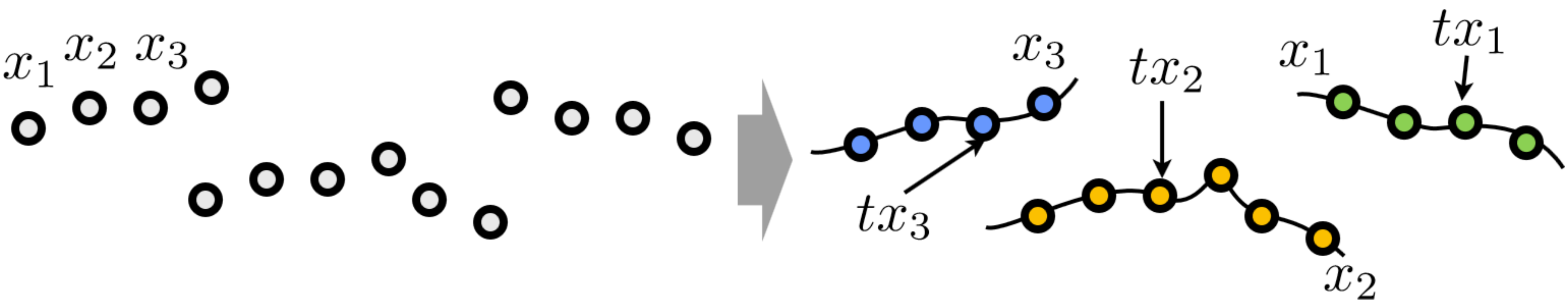
Motion as Natural Transformations

- Consider pedestrian detection in video.
- Training data consists of many sequences of moving persons.
- The frames from the same sequence are highly correlated.
- This breaks the assumption of i.i.d. samples, which is fundamental for most machine learning methods.



Motion as Natural Transformations

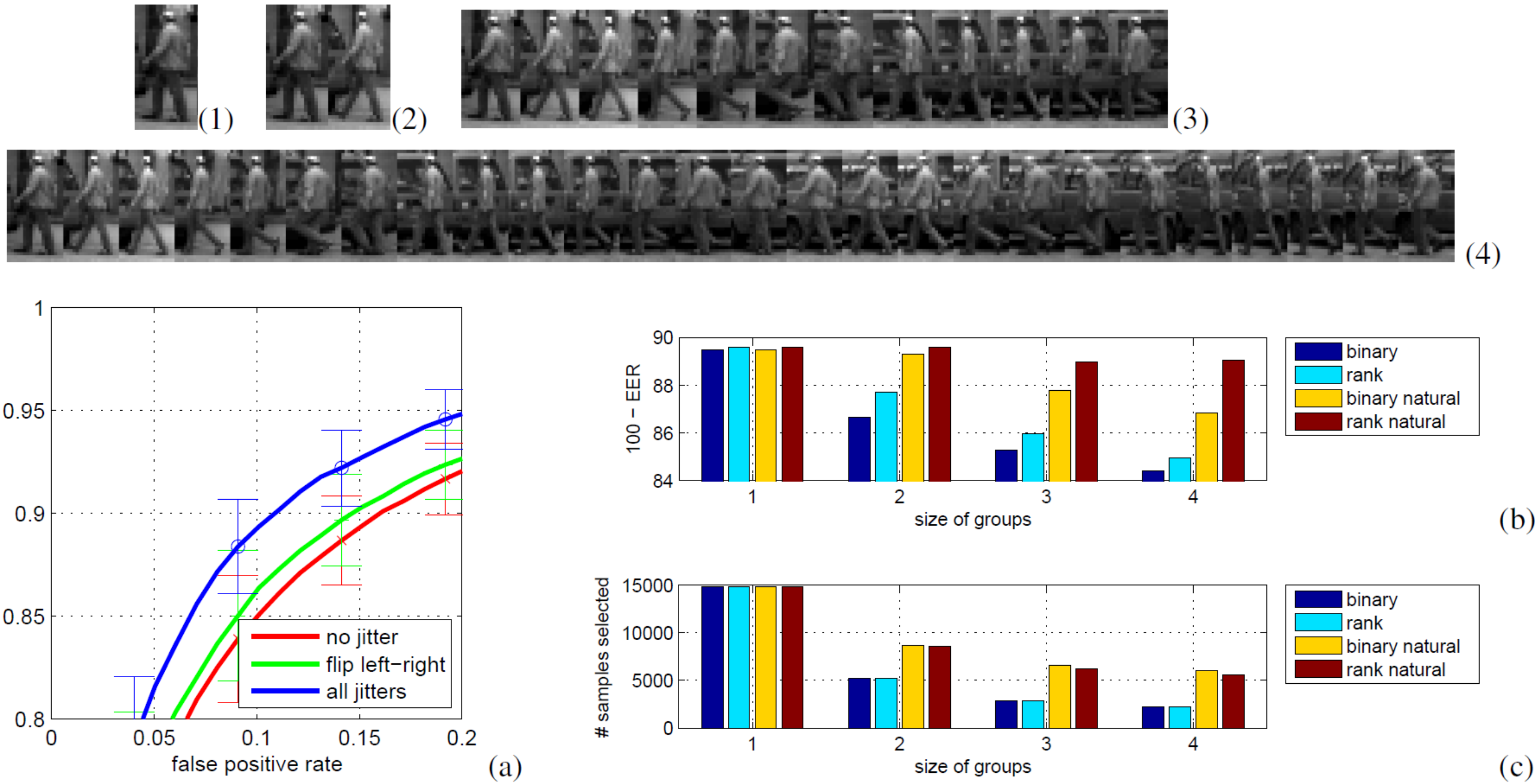
- Consider a sequence as a single training sample, and the different frames in it as transformations of it.



Experiment 2
Pedestrian Classification

- DaimlerChrysler pedestrian classification benchmark. The training data consists of 800 positive images and 5000 negative images, and two test sets of the same size.
- Consider mirroring and translation by 1 pixel as transformations.
- Also consider motion as natural transformations.
- They derive and compare an invariant binary SVM and an invariant rank SVM.

Experiment 2



Conclusion

- The authors propose the use of their algorithm instead of ad-hoc sampling strategies or latent variables to incorporate invariance and equivariance.