Modeling the Impact of Keypoint Detection Errors on Local Descriptor Similarity

André Araujo, H. Lakshman, R. Angst, B. Girod

Department of Electrical Engineering

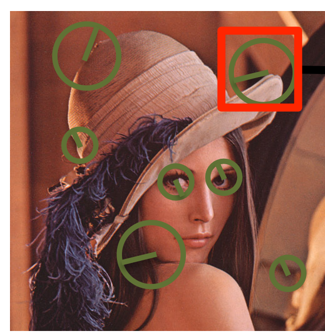
Stanford University

Motivation

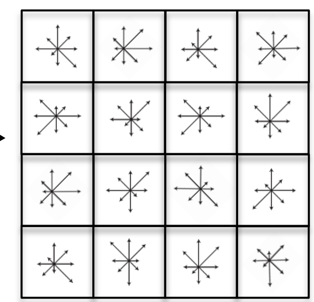
- Gradient-based features: widely used in image processing
  - Motion tracking *[Takacs et al., 2013] [Skrypnyk and Lowe, 2004]*
  - Image-based retrieval *[Duan et al., 2016] [Tao et al., 2014]*
  - Action recognition *[Wang et al., 2013]*
  - Object detection *[Dalal and Trigs, 2005] [Felzenszwalb et al., 2010]*
  - Image classification *[Lazebnik et al., 2006] [Yan et al., 2012]*

- Usual pipeline:

input image



keypoint detection



descriptor extraction using canonical coordinate system

Motivation

- Keypoint detection is sensitive to imaging parameters
  - Empirical studies evaluate robustness of local descriptors to noisy keypoint detection *[Mikolajczyk and Schmid, 2005]*

- Our focus:

  Derive analytical model of local descriptor similarity due to keypoint detection uncertainty

- Several applications:
  - Image retrieval: assess robustness of given descriptor to detection errors
  - Image classification: evaluate grid spacing for dense feature extraction
  - Motion tracking: define required accuracy of a given tracker

Contributions

- First work that models analytically local descriptor similarity as a function of keypoint detection errors

- Main results:

Closed-form expression for L_p distance, for general detection errors

Components of L_2 distance are approximately Gamma-distributed, for translation-only errors

Closed-form expression for expected L_2 distance, for translation-only errors

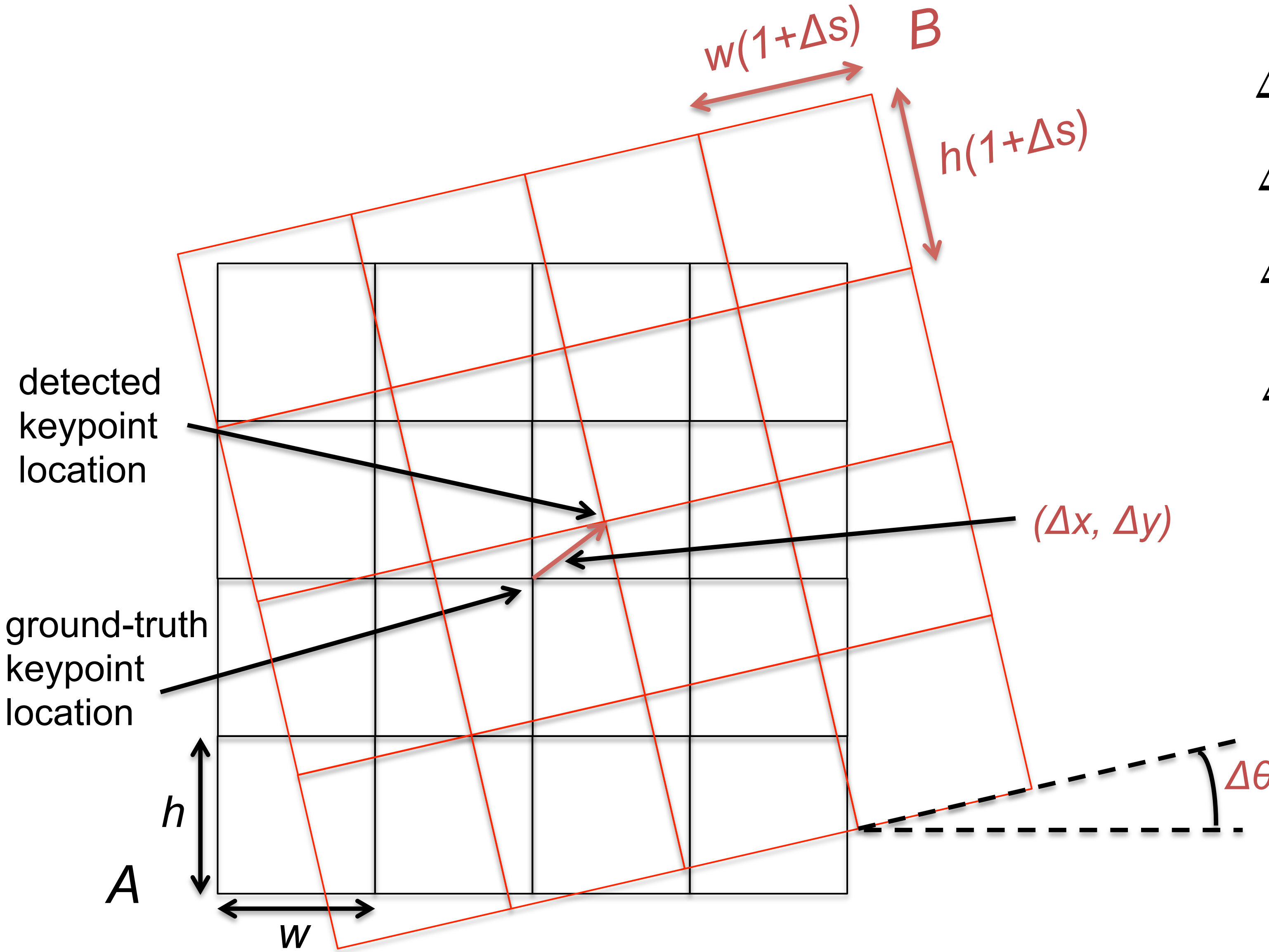
Outline

- Problem Formulation
- General Model
- Detailed Analysis: Translation Errors Only
- Comparison with Experimental Results

Outline

- Problem Formulation
- General Model
- Detailed Analysis: Translation Errors Only
- Comparison with Experimental Results

Problem Formulation



$$\Delta x = x_B - x_A$$

$$\Delta y = y_B - y_A$$

$$\Delta \theta = \theta_B - \theta_A$$

$$\Delta s = \frac{s_B}{s_A} - 1$$

Problem Formulation

- Histogram of gradient orientations:

$$a_n[d]$$

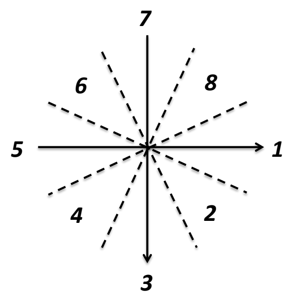
$$\sum_{d=1}^{D} a_n[d] = 1$$ (normalized per spatial bin)

Spatial bins $n = 1, \ldots, N$

Grad. orientation bins $d = 1, \ldots, D$

SIFT: *N=16*

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16



SIFT: *D=8*

SIFT: *128*-dim

- Local descriptor: $f_A = [a_1[1], a_1[2], \ldots, a_1[D], \ldots, a_N[D]]$

- We are interested in modeling: $\| f_A - f_B \|_p^p$

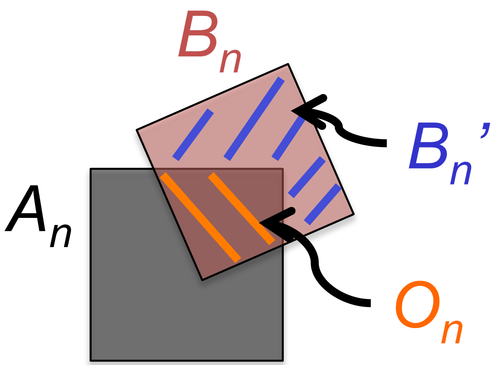
Outline

- Problem Formulation
- General Model
- Detailed Analysis: Translation Errors Only
- Comparison with Experimental Results

Main result:

Closed-form expression for L_p distance, for general detection errors

General Model



- A_n *n*-th spatial bin of patch A
- B_n *n*-th spatial bin of patch B
- O_n overlap region of A_n and B_n
- B_n' non-overlap region of B_n

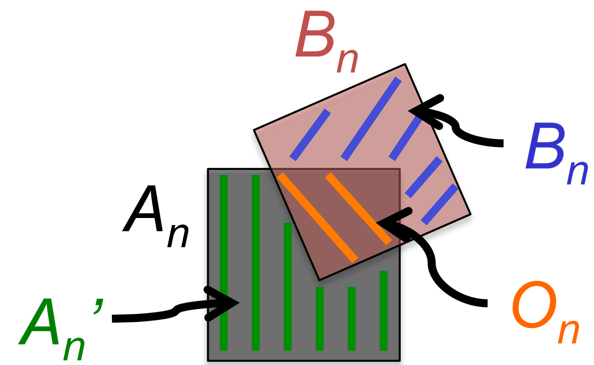
Normalized histogram of B_n as a function of those from O_n, B_n'

$$b_n[d] = \beta_n o_n^B[d] + (1 - \beta_n) b_n'[d]$$

Proportion of overlap and non-overlap areas (sum to 1)

General Model: Descriptor Distance

Writing out a similar expression for A_n and rearranging terms, we obtain:



$$\|f_A - f_B\|_p^p =$$

$$\sum_{n=1}^{N}\sum_{d=1}^{D}\left|\boxed{(1-\beta_n)(a'_n[d]-b'_n[d])}\right.$$ → compares A_n' and B_n'

$$+\boxed{\beta_n(o_n^A[d]-o_n^B[d])}$$ → compares O_n (with A's and B's references)

$$\left.+\boxed{\beta_n(2\Delta s+\Delta s^2)(o_n^A[d]-a'_n[d])}\right|^p$$ → compares histograms of O_n and A_n'

Outline

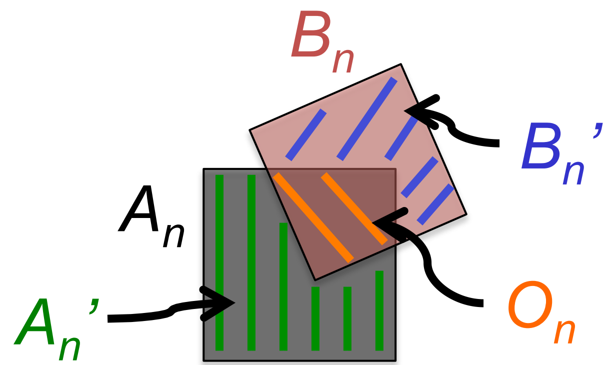
- Problem Formulation
- General Model
- Detailed Analysis: Translation Errors Only
- Comparison with Experimental Results

Main result:

Closed-form expression for expected L_2 distance, for translation-only errors

Translation Errors Only: Simplification

General expression from before:



$$\| f_A - f_B \|_p^p =$$

$$\sum_{n=1}^{N} \sum_{d=1}^{D} \Big| \boxed{(1-\beta_n)(a'_n[d] - b'_n[d])}$$

→ compares A_n' and B_n'

$$+ \boxed{\beta_n (o_n^A[d] - o_n^B[d])}$$

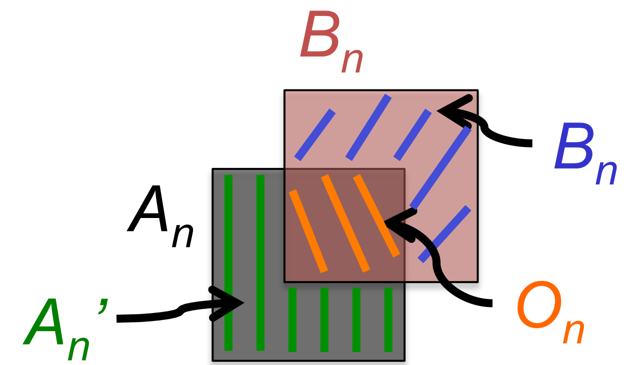
→ compares O_n (with A's and B's references)

$$+ \boxed{\beta_n (2\Delta s + \Delta s^2)(o_n^A[d] - a'_n[d])} \Big|^p$$

→ compares histograms of O_n and A_n'

Translation Errors Only: Simplification

In this case: $\Delta\theta \approx 0$, $\Delta s \approx 0$



$$\| f_A - f_B \|_p^p =$$

$$\sum_{n=1}^{N} \sum_{d=1}^{D} \left| \boxed{(1-\beta_n)(a'_n[d] - b'_n[d])} \right.$$

→ compares A_n' and B_n'

$$+ \cancelto{0}{\boxed{\beta_n (o_n^A[d] - o_n^B[d])}}$$

→ ~~compares O_n (with A's and B's references)~~

$$\left. + \cancelto{0}{\boxed{\beta_n (2\Delta s + \Delta s^2)(o_n^A[d] - a'_n[d])}} \right|^p$$

→ ~~compares histograms of O_n and A_n'~~

Translation Errors Only: Expected Value

Using p=2:

$$\|f_A - f_B\|_2^2 = \sum_{n=1}^{N} \sum_{d=1}^{D} |(1-\beta_n)(a'_n[d] - b'_n[d])|^2$$

We are interested in estimating the mean of such descriptor distance:

$$E[\|f_A - f_B\|_2^2] = \sum_{n=1}^{N} \sum_{d=1}^{D} \boxed{E[|(1-\beta_n)(a'_n[d] - b'_n[d])|^2]}$$

key term

Expressing Histogram Using Binary Masks

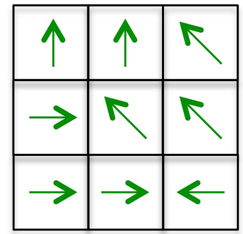
Define the binary mask:

$$g_d[x,y]$$

d: orientation bin

[x,y]: position

Gradients



$g_1[x,y]$

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

→ corresponds to $d=1$

$$a_n[1] = \frac{1}{3}$$

We can write:

$$a_n[d] = \boxed{\frac{1}{\#\text{pixels}}} \boxed{\sum_{x,y} g_d[x,y]}$$

Normalize by number of pixels in region

Number of pixels with gradient quantized to *d*

Assumptions

Assumption 1: $a'_n[d]$ and $b'_n[d]$ are uncorrelated and identically distributed

Assumption 2: statistics of $g_d[x,y]$

- Option 1 (**Strong**): $g_d[x,y]$ is IID $\longrightarrow$ M-IID
- Option 2 (**Mild**): $g_d[x,y]$ is stationary $\rightarrow$ M-S

Models for Different Scenarios

- Fixed translation errors

  Obtained by using derivations and assumptions from previous slides

- Uniformly-distributed translation errors

  Use iterated expectation, given results with fixed translation errors

- In both cases, we obtain closed-form expressions

Outline

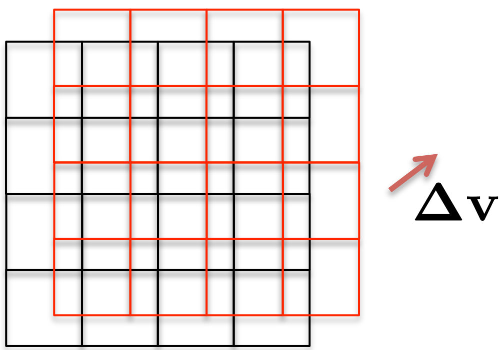
- Problem Formulation
- General Model
- Detailed Analysis: Translation Errors Only
- Comparison with Experimental Results

Experimental Setup

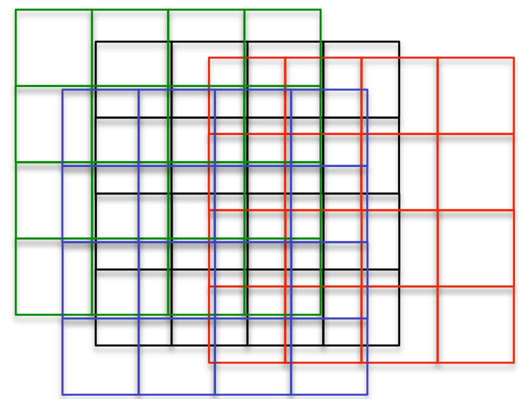
- Two datasets, with two different keypoint detectors
  - Stanford Mobile Visual Search (SMVS) dataset *[Chandrasekhar et al., 2011]*
    - 65k keypoints extracted with DoG detector (as in SIFT)
  - CNN2h dataset *[Araujo et al., 2014]*
    - 78k keypoints extracted with TCD detector *[Makar et al., 2014]*
  - Datasets divided into train/test splits
- 4x4 spatial bins, 8 gradient orientations (as in SIFT)

Experimental Setup

- Experiments with fixed translation errors $\mathbf{\Delta v}$



- Experiments with uniform translation errors
$$-\frac{U}{2} \leq \mathbf{\Delta v} \leq \frac{U}{2}$$



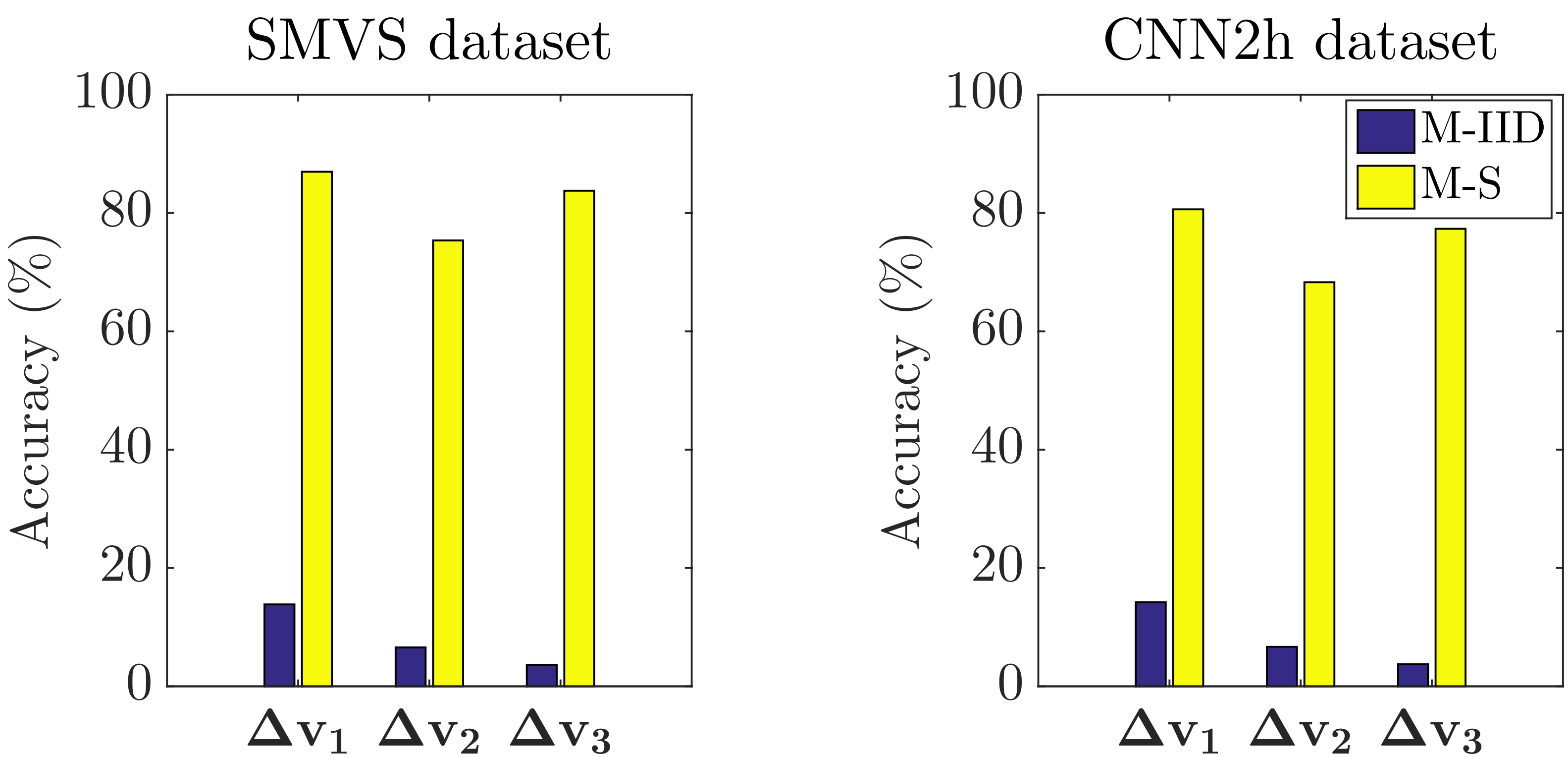
- We compare empirical versus estimated expected values of descriptor distances
- Accuracy of estimates given by: *Acc = 1 – RelativeError* (higher is better)

Experiments: Fixed Translation Error

- We use three different translations:

$$\mathbf{\Delta v_1} = [1, 1]$$
$$\mathbf{\Delta v_2} = [-1, 3]$$
$$\mathbf{\Delta v_3} = [4, -4]$$

- Results:

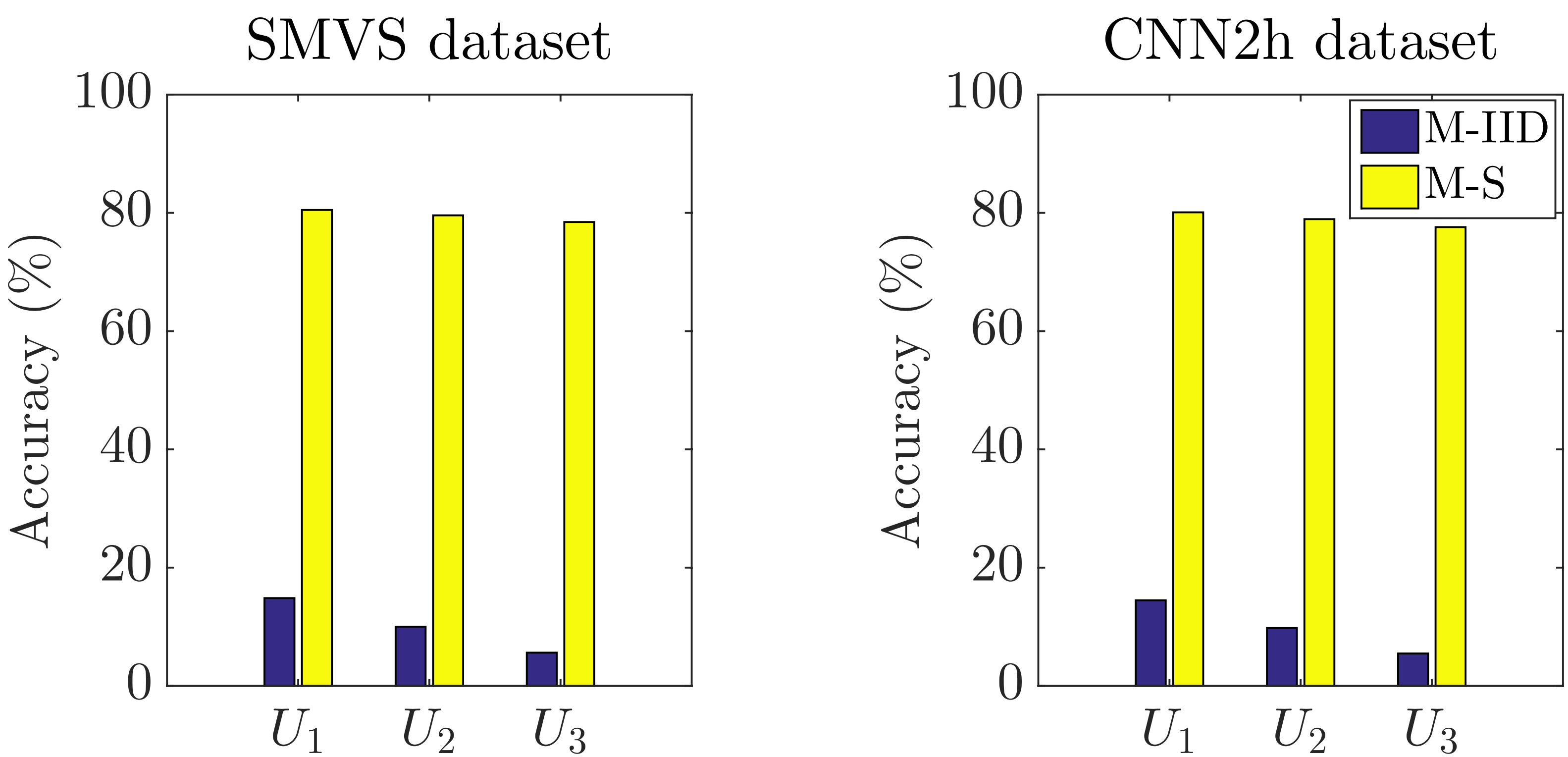


Experiments: Uniform Translation Errors

- We use three different distributions: $-\frac{U}{2} \leq \mathbf{\Delta v} \leq \frac{U}{2}$

  with $U_1 = 2 \quad U_2 = 4 \quad U_3 = 8$

- Results:



Conclusions

- First work to model analytically descriptor similarity as a function of keypoint detection errors
- We develop expression for L_p distance based on general translation, orientation and scale detection errors
- Proposed stationary model explains most of the variation of descriptor distance when translation errors dominate
- Framework can be modified to analyze other binning configurations

Thank you! Questions?

André Araujo

http://stanford.edu/~afaraujo

afaraujo@stanford.edu