NTT

トピックモデルの応用：画像・動画像データ

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

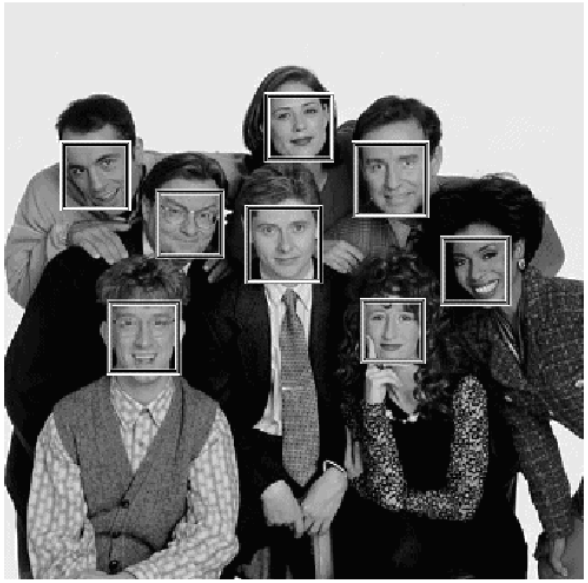
このスライドの“トピック”

- （動）画像認識・処理(Computer Vision)は最先端の機械学習技術が素早く導入される分野です

- この分野でもトピックモデルは猛威を振るっています（いました？）

Computer Vision（機械視覚・・・？）

- 人工知能の黎明期から、機械学習・パターン認識のもっとも分かり易い応用分野です



[Viola & Jones, 2001]

[Shimizu, 2008]

機械学習とcv

- 最先端のパターン認識・機械学習研究との親和性が非常に高いです
  - 大きなデータサイズ、リッチなコンテキスト
  - 実世界の複雑さ
  - 色々なことが可能

すべてを概観することは不可能なので

- 今回はトピックモデル応用がある論文まわりだけです

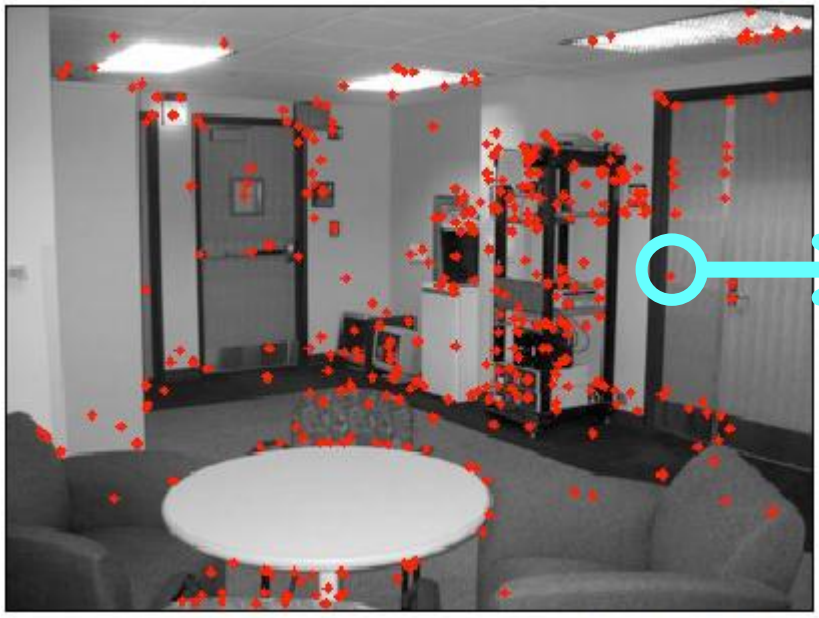
- その前に、あらゆる手法でほぼ共通して利用される基本的なアイデアを説明します

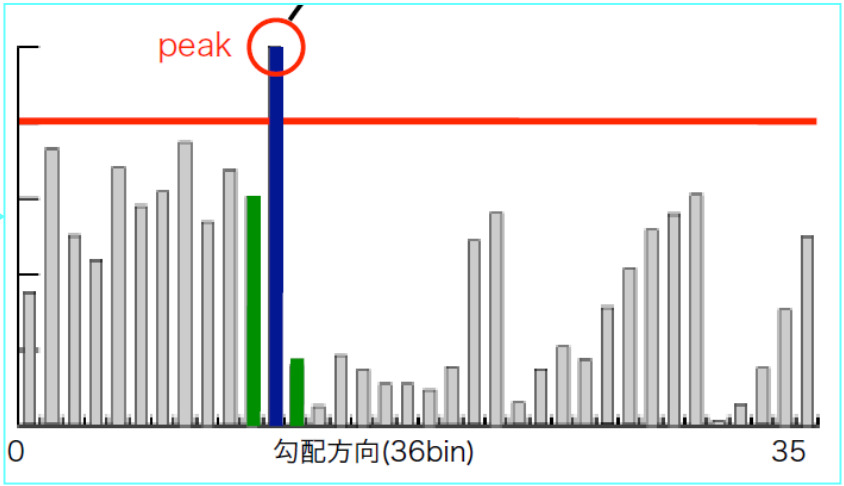
SIFT [Lowe, 2004]

- 画像認識系研究のデファクトスタンダードな特徴量

被引用数: 16268 (as of 2012/10/30, Google Scholar)

- 画素値勾配変化の局所極大・極小点を検出
- 注目点付近の画像勾配分布を計算

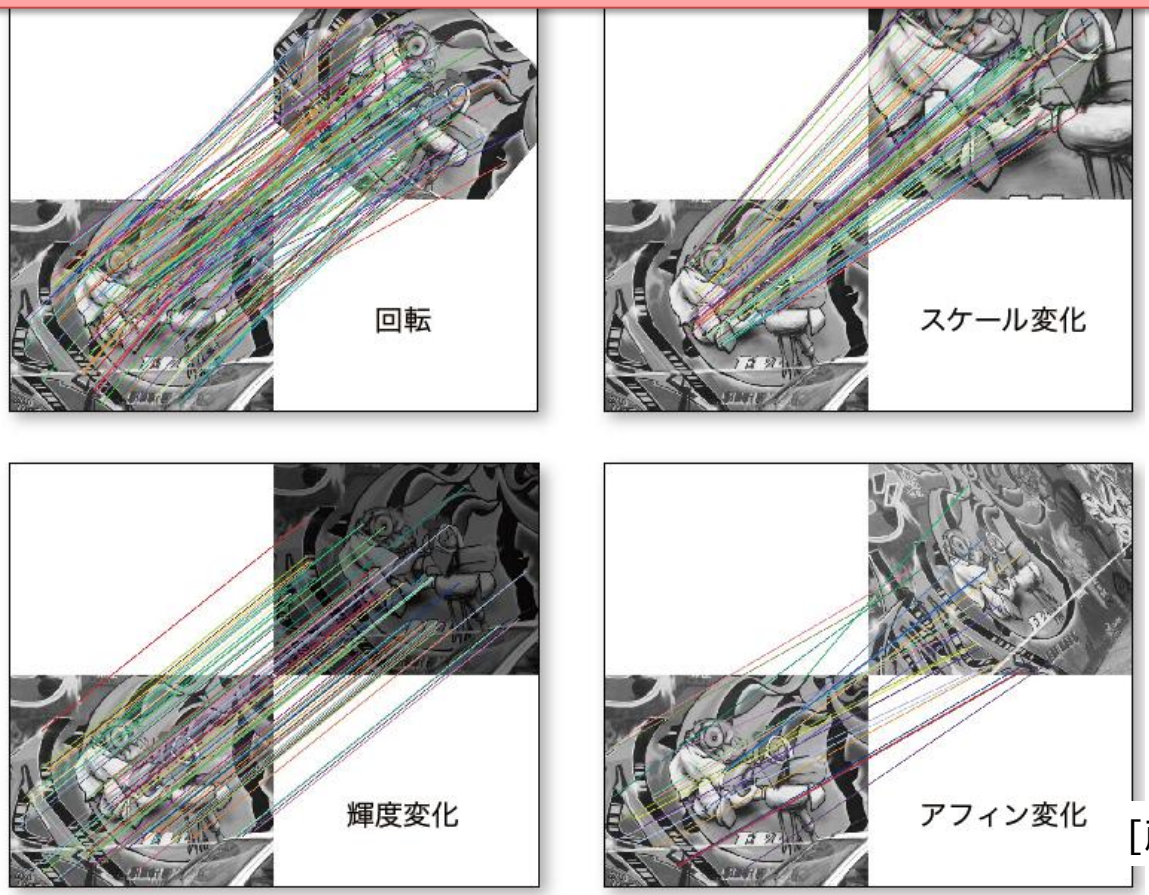




[藤吉, 2007]

- 画像の回転･スケール変化に対し不変
- 照明変化に対してロバスト

画像間の対応点検出に理想的な性質
→画像のパノラマ合成、物体認識、画像分類など
ありとあらゆる認識系のタスクで利用されている☺



[藤吉, 2007]

SIFT [Lowe, 2004]

- 近年は高速化・特徴圧縮した拡張法・関連法もよく利用される
- ☺ 中部大学・藤吉先生による日本語チュートリアルをお勧めします

藤吉, "Gradientベースの特徴抽出 - SIFTとHOG - ",
情報処理学会 研究報告 CVIM 160, pp. 211-224, 2007.

(参考文献にも載せてあります)

Bag of Visual Words: 画像データの「文書化」

- 局所画像特徴量(含むSIFT)は1枚の画像からたくさん計算できます

- 1つ1つの特徴量は高次元ベクトルです
  - (SIFTだと128次元実数ベクトル)

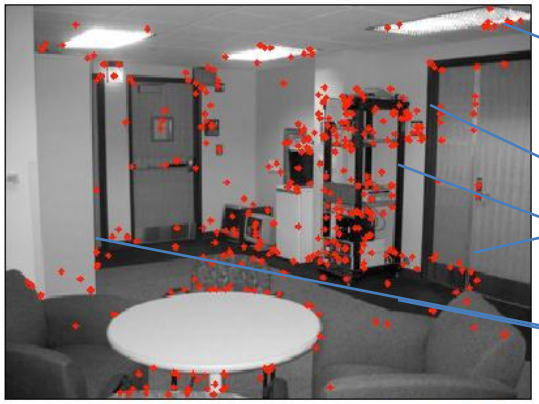
- そこで、適当にサボった表現がほしくなります
  → Bag Of Visual Words

Bag of Visual Words: 画像データの「文書化」

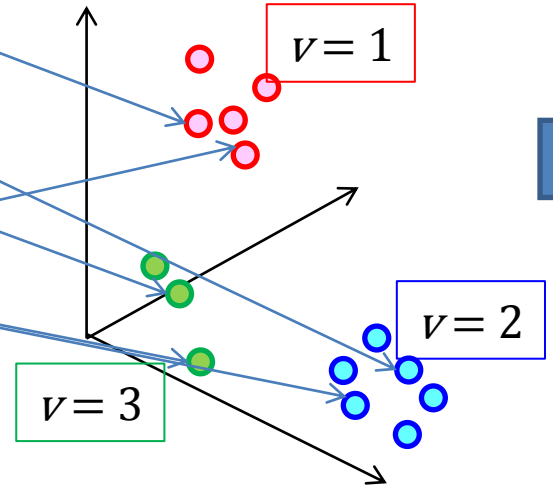
局所特徴を抽出

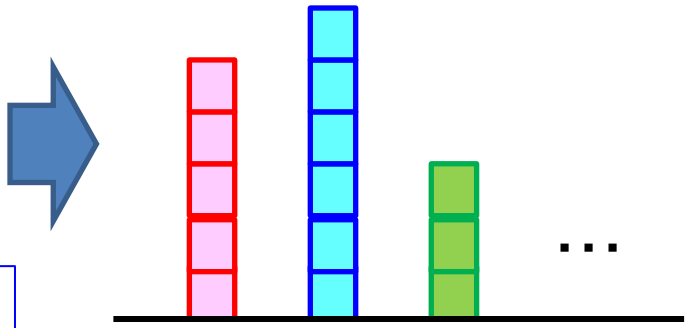
K-meansなどによる量子化

V次元のヒストグラム



[藤吉, 2007]





Visual Words:
単語に相当

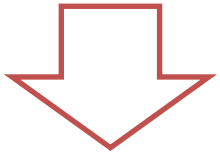
Bag of Visual Words:
文書に相当

☺ 機械学習・自然言語処理で開発された
各種学習モデルをそのまま利用できる
☺ メモリ・計算量削減

SIFT + Bag of Visual Words = 安心

- ここ数年は、SIFTをBoVWで量子化したものを観測特徴量として利用する研究ばかりです

- SIFT（あるいはその進化系）: 高性能な基本特徴量
- Bo(V)W: 計算量削減、機械学習技術との連携が容易



画像データでもトピックモデル！！

Scene Recognition
[Fei-Fei and Perona, 2005]

Fei-Fei and Perona,
"A Bayesian Hierarchical Model for Learning
Natural Scene Categories",
in Proc. CVPR. 2005.

“mountain”

“ocean”

“forest”

“ocean mountain”

人間はどのように sceneを認識しているのか？

- 画像中の全ての物体・背景を個別に認識して、それらから推測されるコンテキストとしてsceneを認識する [Treisman & Gelade, 1980]

- No!!

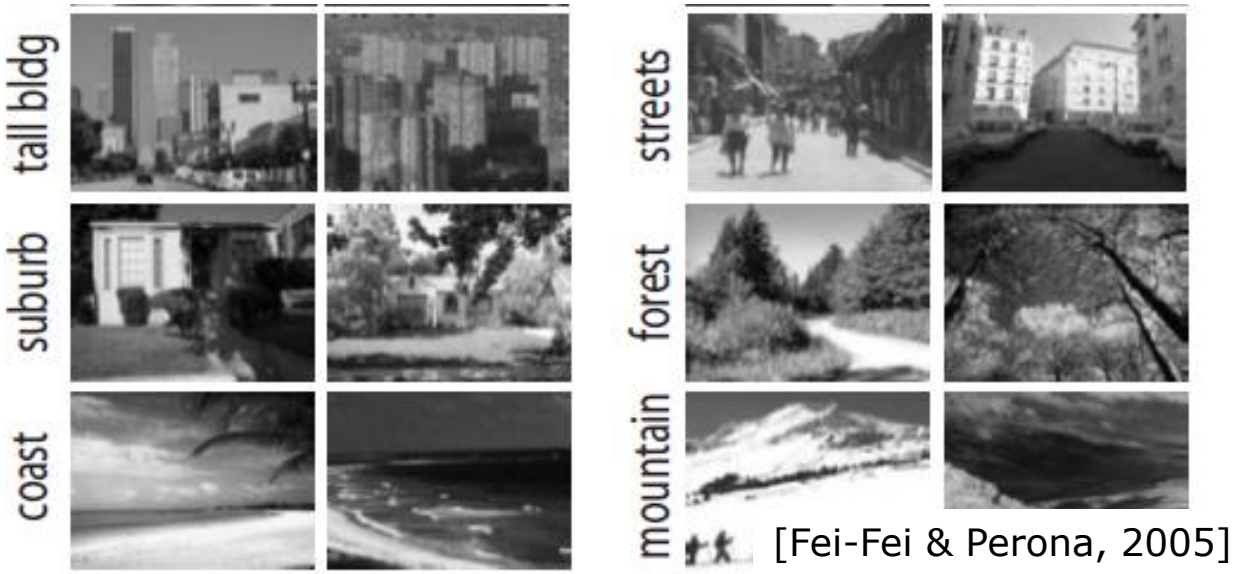
- 人間はほとんど注視をすることなく、画像のsceneを識別できる[Li, 2002]

Computer visionでも そうですか？

- 色々な研究者がいろいろ試しました
- 結論：生の画像情報だけでなく、それを「意味のある良いパターン」に抽象化した表現が作れるといい性能が得られる
- 問題：「良いパターン」を人間が設計するのは非常に大変

提案法: Theme model for scene recognition

- Scene recognitionにおいて、「意味のあるパターン＝theme」を使うと良い
- トピック＝themeと仮定してトピックモデルを適用してみた→うまくいった！☺

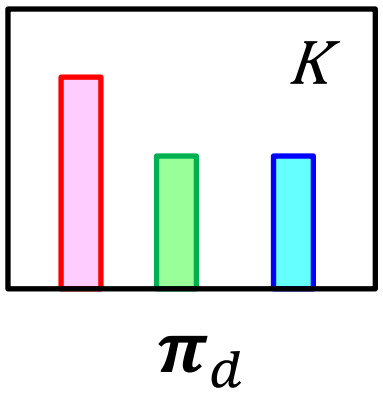


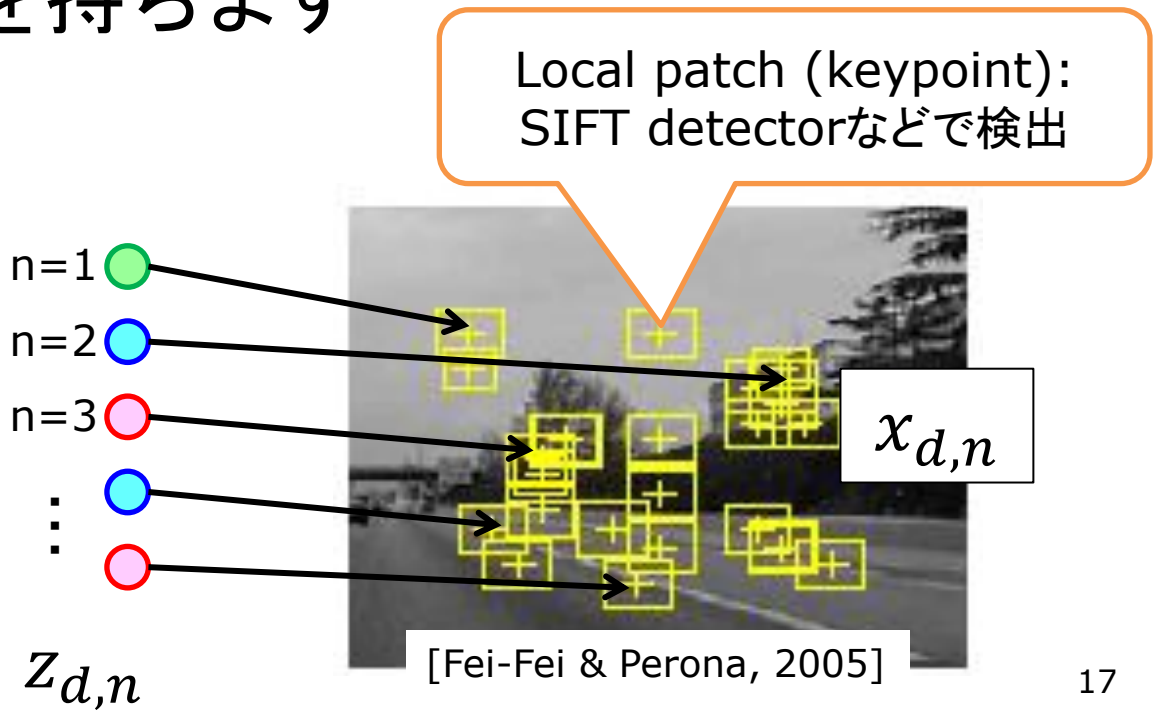
[Fei-Fei & Perona, 2005]

提案法のアイデア: 画像の表現

- 画像が文書で、単語は局所的なパッチ（小画像）のVisual Wordです
- 各画像ごとに特有のトピック分布、つまりパッチのパターンを持ちます

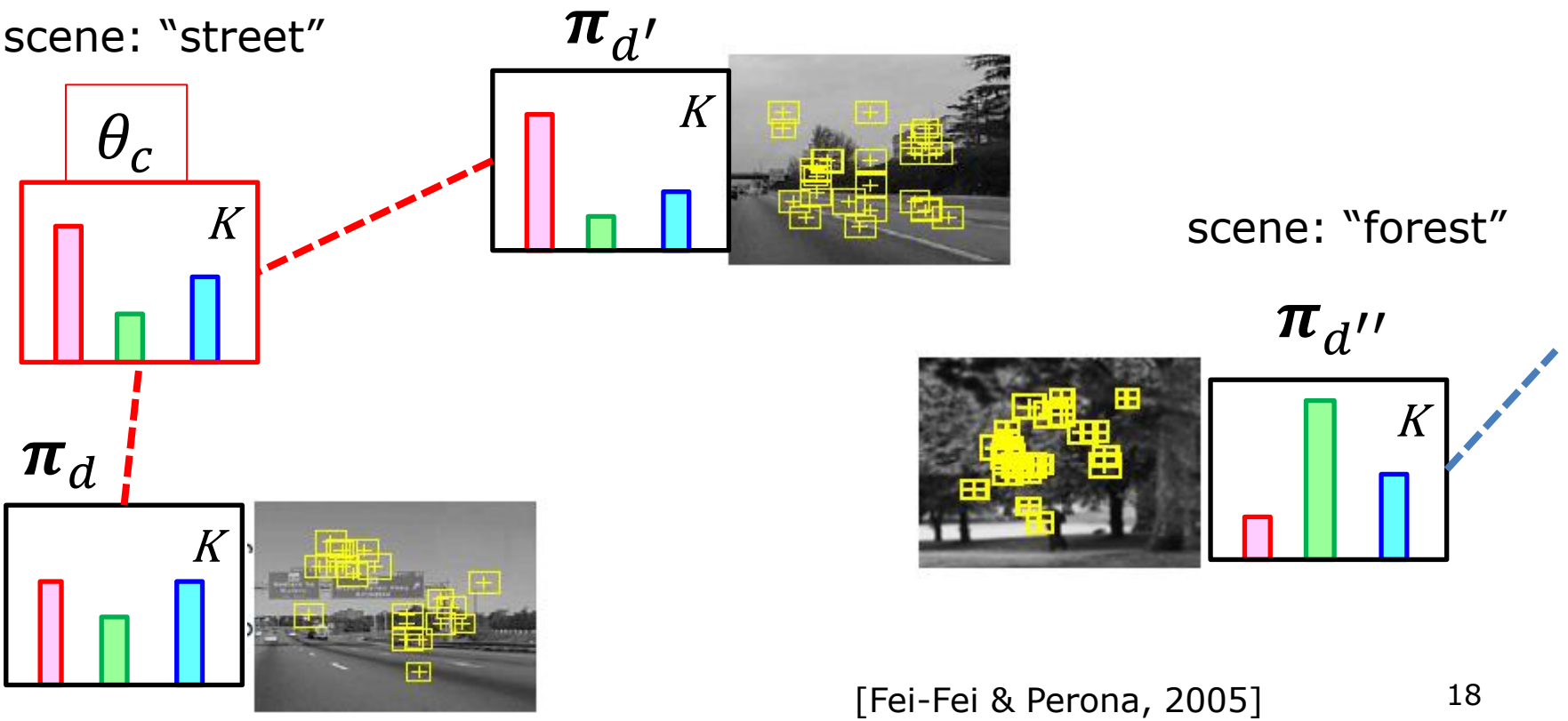
画像(文書) d





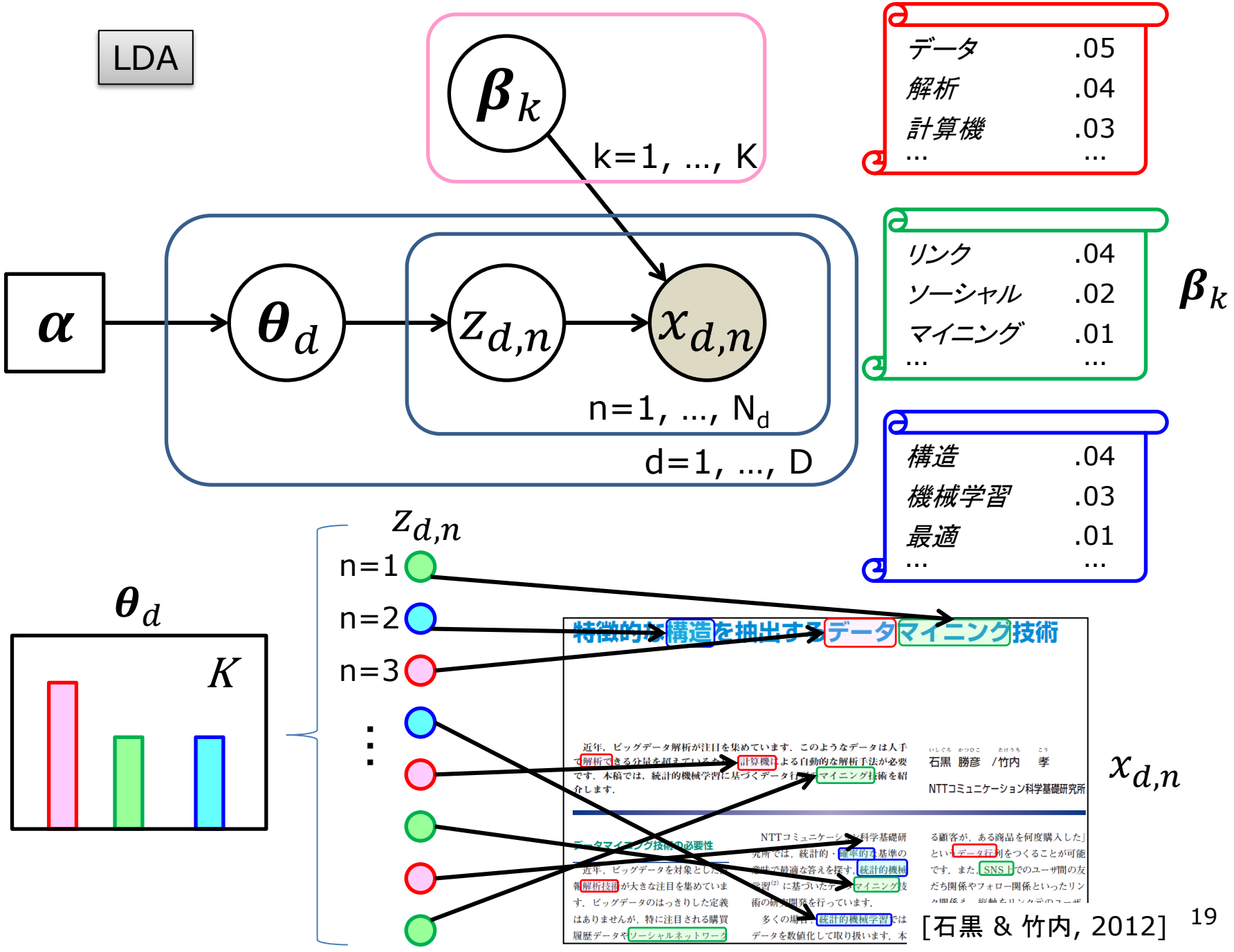
提案法のアイデア: sceneカテゴリのトピック中心

- sceneとしてカテゴライズできる以上、同じsceneカテゴリの画像は相関があるはず



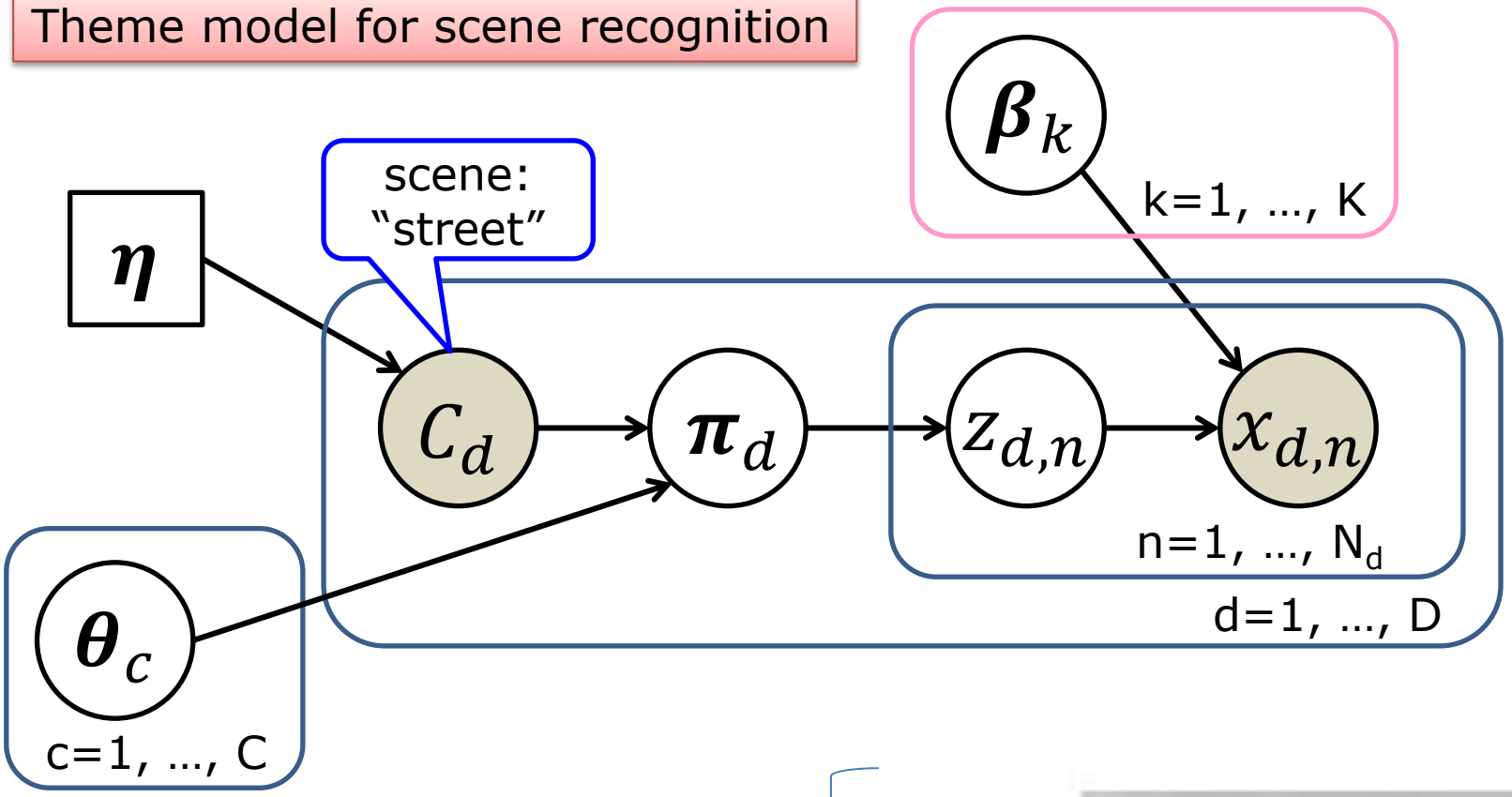
[Fei-Fei & Perona, 2005]

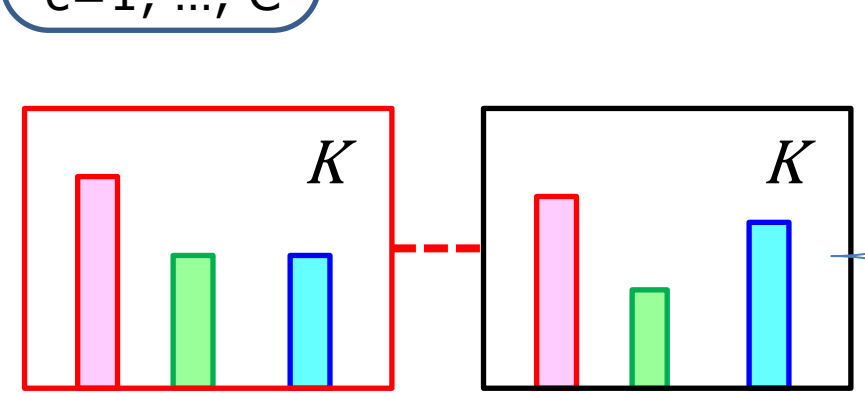
LDA

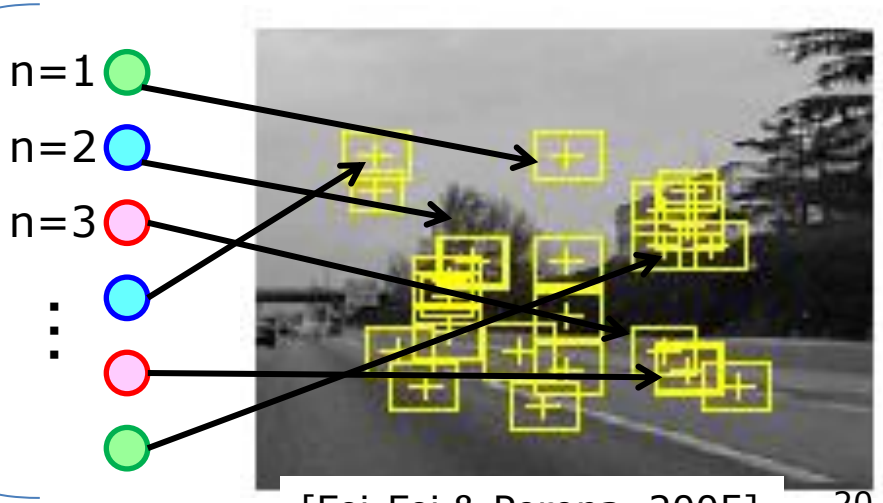


[石黒 & 竹内, 2012]

Theme model for scene recognition







[Fei-Fei & Perona, 2005]

生成モデル

for 画像d = 1, 2, …, D

scene category label $c_d|\boldsymbol{\eta} \sim \mathrm{Mult}(\boldsymbol{\eta})$

topic proportion $\boldsymbol{\pi}_d|\boldsymbol{\theta}_{c_d} \sim \mathrm{Dir}(\boldsymbol{\theta}_{c_d})$

for 単語n = 1, 2, …, N_d

topic-VW assignment $z_{d,n}|\boldsymbol{\pi}_d \sim \mathrm{Mult}(\boldsymbol{\pi}_d)$

VW observation $x_{d,n}|z_{d,n}, \{\boldsymbol{\beta}_k\} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$

for theme (topic) k = 1, 2, …, K

topic-VW proportion $\boldsymbol{\beta}_k$

for sceneカテゴリ c = 1, 2, …, C

“average” topic proportion $\boldsymbol{\theta}_c$

パラメータ、隠れ変数の推定

- 論文では変分ベイズ(VB)による推定法が紹介されています

- 通常のLDAの解法を参考にすれば、解は比較的簡単に導出できるようなので、ここでは割愛します

未知画像の識別

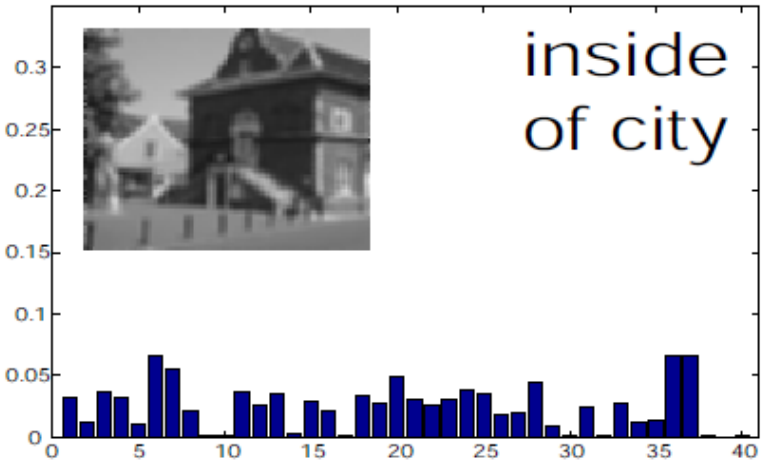
- 学習し終わったモデルに対して、未知画像 d のsceneカテゴリ c を推定します
- 最尤推定で計算しますが、正確な計算は不可能です
  - 適切な近似が必要ですが、これもLDAの元論文等を参考にしてください

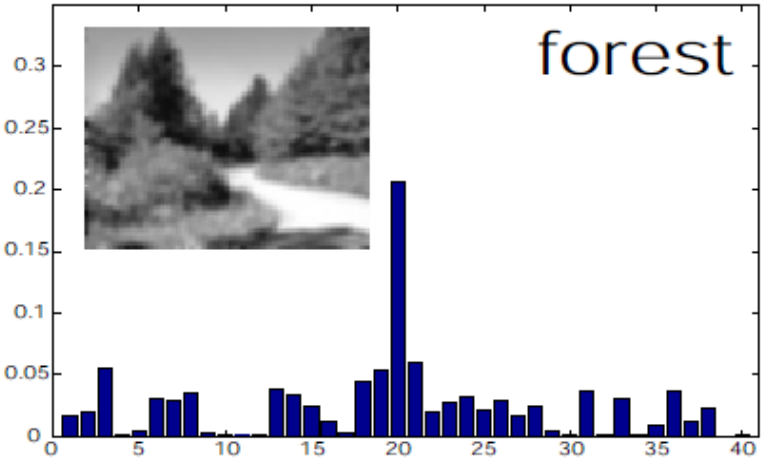
$$c_d = \arg\max_c p(X_d|c, \boldsymbol{\theta}, \boldsymbol{\beta})$$

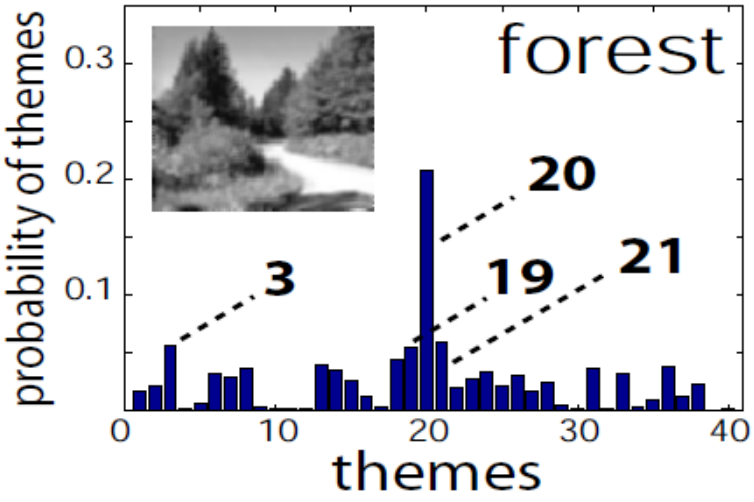
$$= \int p(\boldsymbol{\pi}_d|c, \boldsymbol{\theta}) \left\{ \prod_n \sum_k p(z_{d,n} = k|\boldsymbol{\pi}_d) p(x_{d,n}|z_{d,n}, \boldsymbol{\beta}) \right\} d\boldsymbol{\pi}_d$$

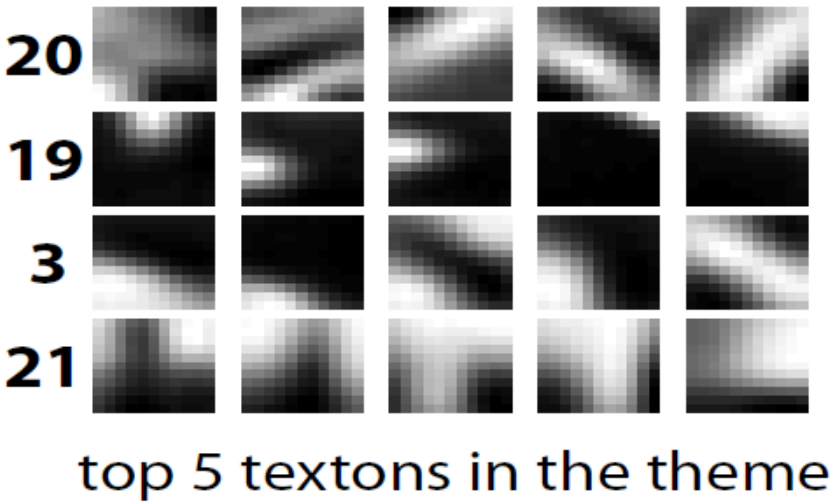
	# of categ.	training # per categ.	training requirements	perf. (%)
Theme Model 1	13	100	unsupervised	76
[17]	6	$\sim$ 100	human annotation of 9 semantic concepts for 60, 000 patches	77
[9]	8	250 $\sim$ 300	human annotation of 6 properties for thousands of scenes	89

[Fei-Fei & Perona, 2005]









[Fei-Fei & Perona, 2005]

まとめ: Theme model for Scene Recognition

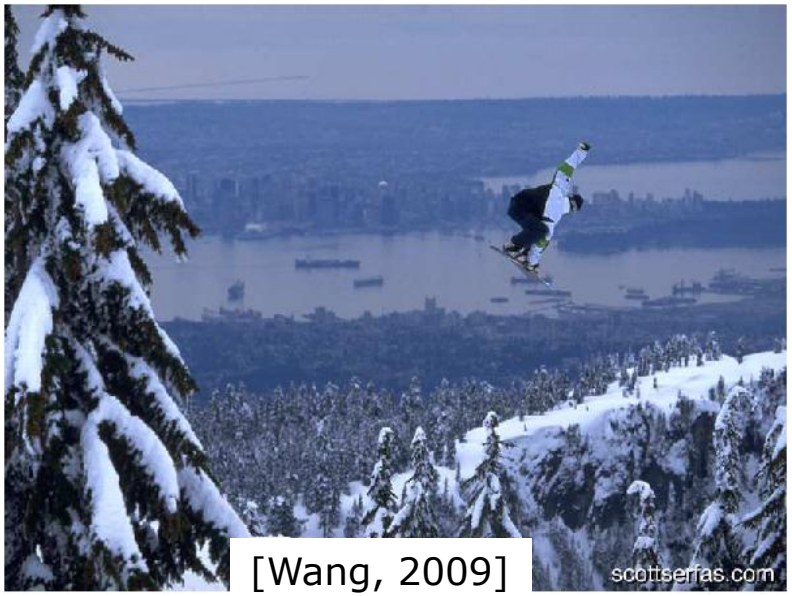
- BoVWを素直に使った（だけ）の、トピックモデル応用
- 13カテゴリの識別問題に対して、事前にパッチやthemeを人手で与える必要がない
- Theme (topic) に意味があるかは・・・？？

Scene Classification with Annotation [Wang, 2009]

Wang, Blei and Fei-Fei,
"Simultaneous al Model for Learning
Natural Scene Categories",
in Proc. CVPR. 2009.

画像カテゴリの識別問題 (image classification)

- 「この画像は何か当てて下さい」
- 画像に対して、有限のクラスのうち最も適切なものを当てる問題
- 古くからの画像認識問題



[Wang, 2009]

“running”

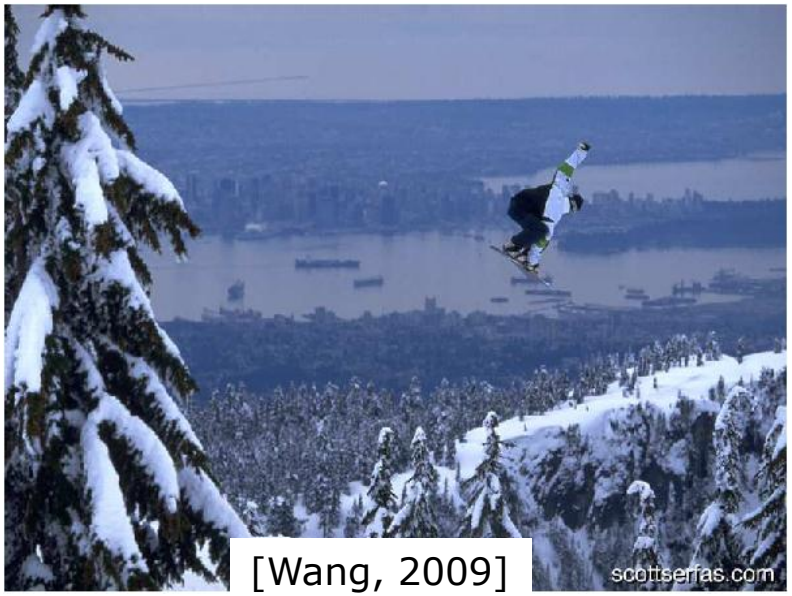
“skiing”

“snowboarding”

“painting”

画像のアノテーション付与問題 (image annotation)

- 「この画像に適切なタグを付与してください」
- 画像に対して、関連するタグ（単語）を複数付与する問題
- 画像検索の文脈などで幅広く研究



[Wang, 2009]

○ “forest”

○ “snowboarder”

○ “sky”

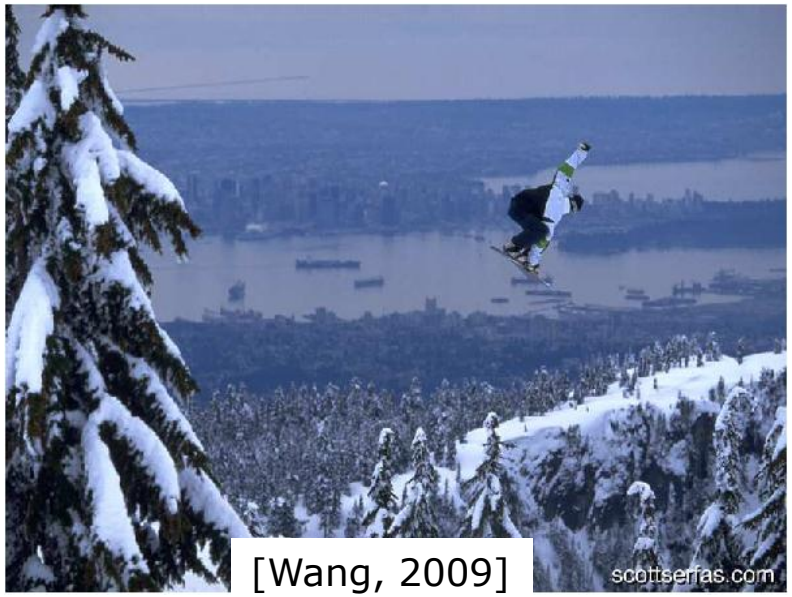
○ “jump”

× “track”

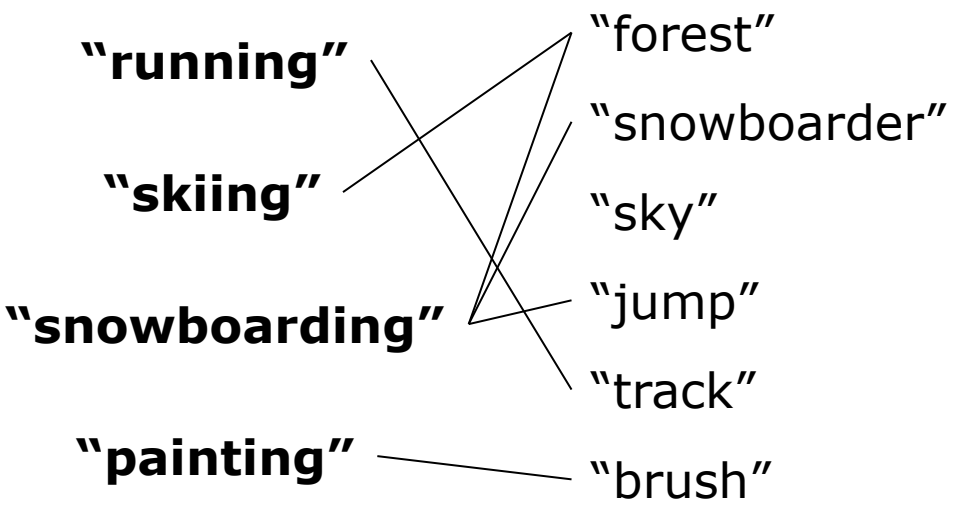
× “brush”

カテゴリ識別とアノテーション、実はすごく近い問題？

- クラスラベルとタグは間違いなく相関がある
- 両方を用いることで、より高精度に両問題を解ける？



[Wang, 2009]



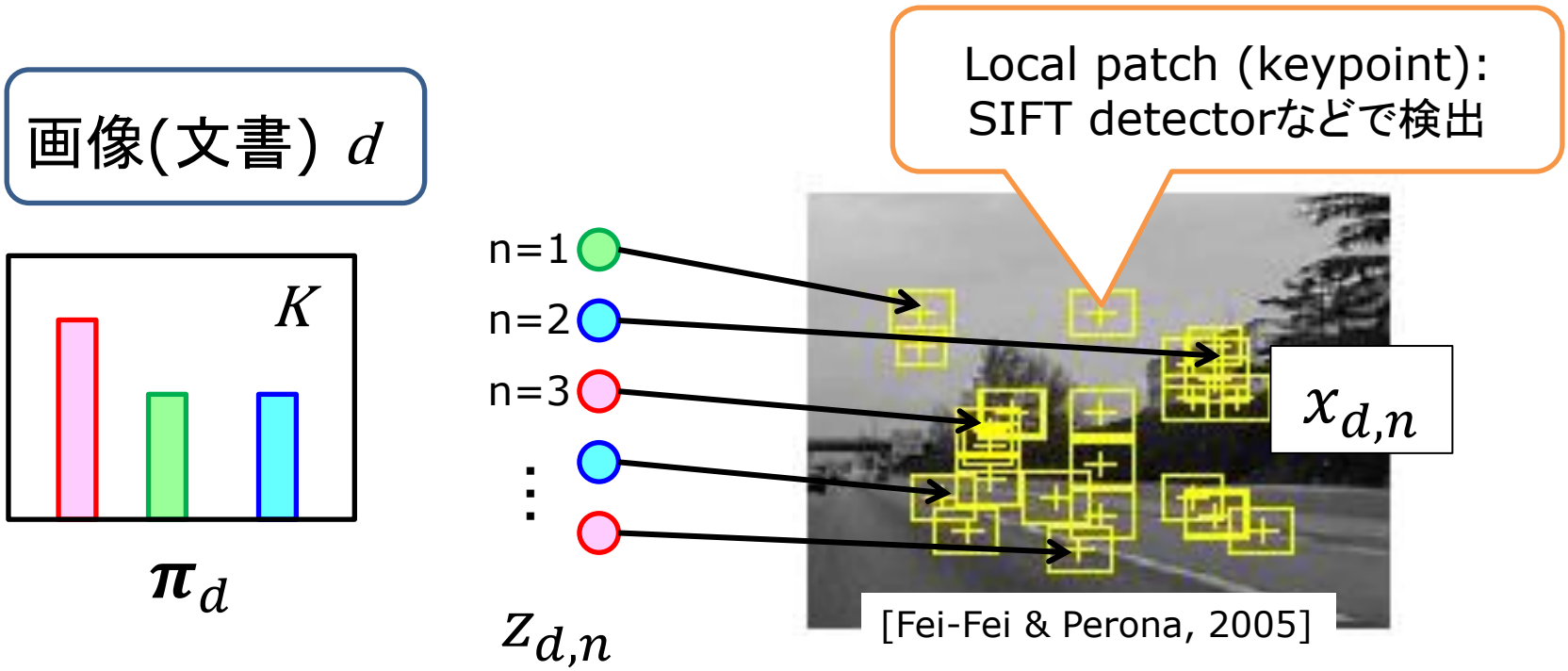
提案法: Multi-class sLDA with annotation

- 画像のカテゴリ識別(classification)とタグ当て(annotation)を、一つのトピックモデルで同時にモデル化

- 単一のモデルよりも両課題について高精度
- annotation無しでも、多クラス識別可能なトピックモデルとして新規性・応用性が高い

提案法のアイデア：画像の表現

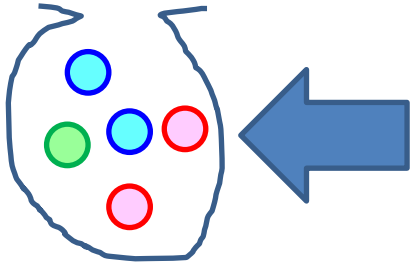
- 先ほどのモデルと同じです
- Local patchがN_d個あります



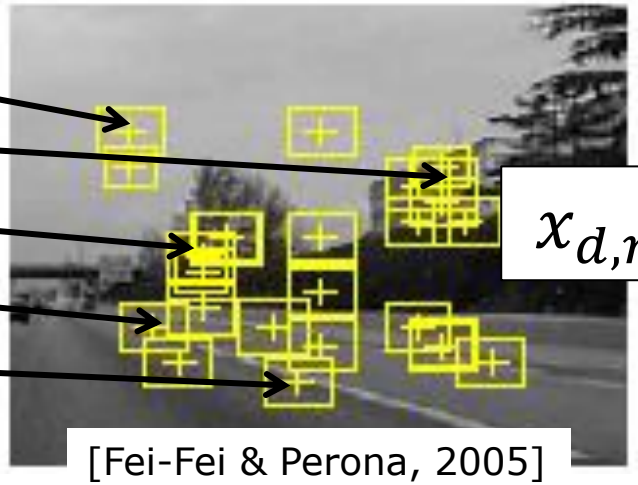
提案法のアイデア：画像とクラスラベルの関係

- 画像のクラスラベルは、画像全体のトピック割り当てから決定します

$$\bar{z}_d = \frac{1}{N_d}\sum_n z_{d,n}$$



n=1
n=2
n=3
⋮



$x_{d,n}$

[Fei-Fei & Perona, 2005]

Soft-max

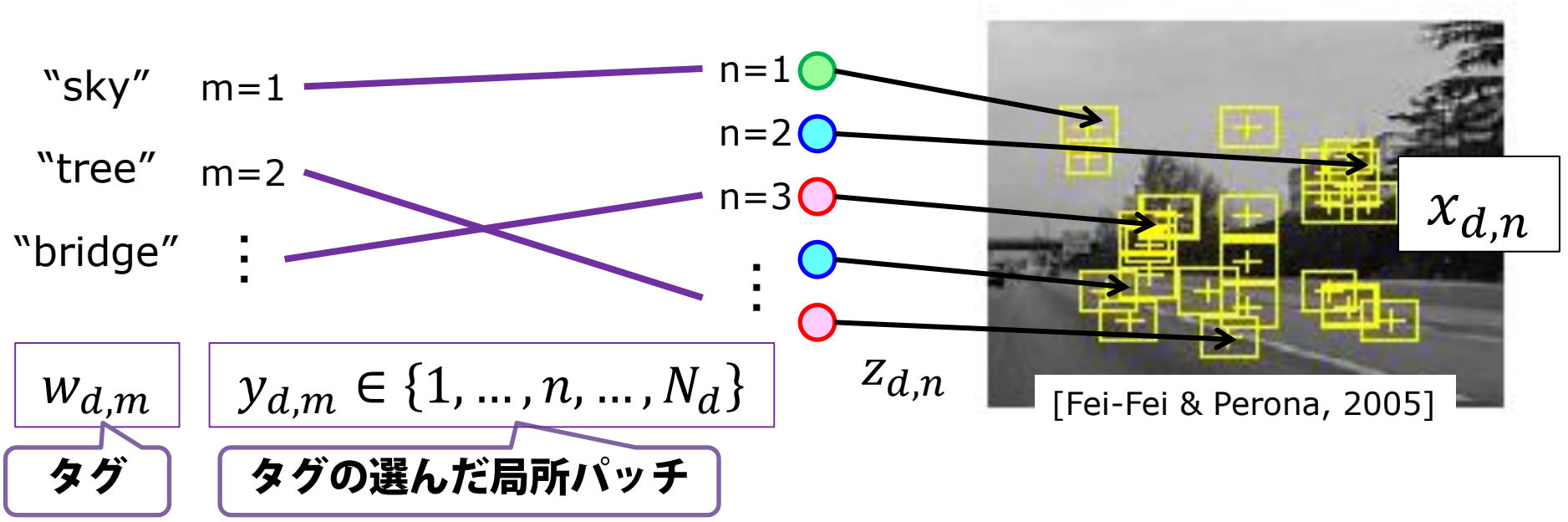
$z_{d,n}$

$c_d =$ “mountain” “water” “city” “street”

画像 d のクラスラベル

提案法のアイデア：画像とタグの関係

- 画像中の局所パッチが、あるタグの「生成元」になっていると考えて、トピックも共有します



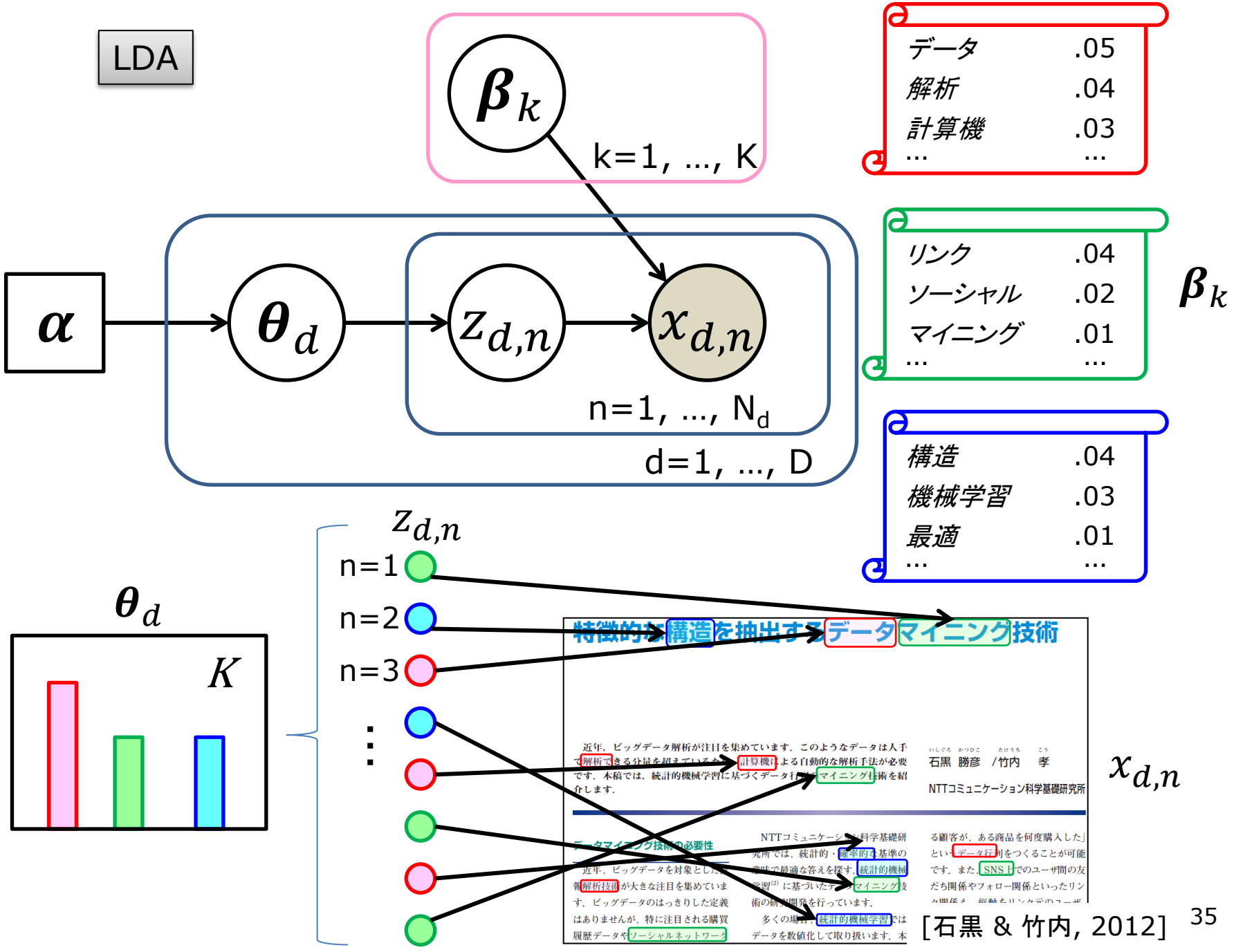
β_k

highway	.04
bridge	.03
…	…

sky	.04
cloud	.02
…	…

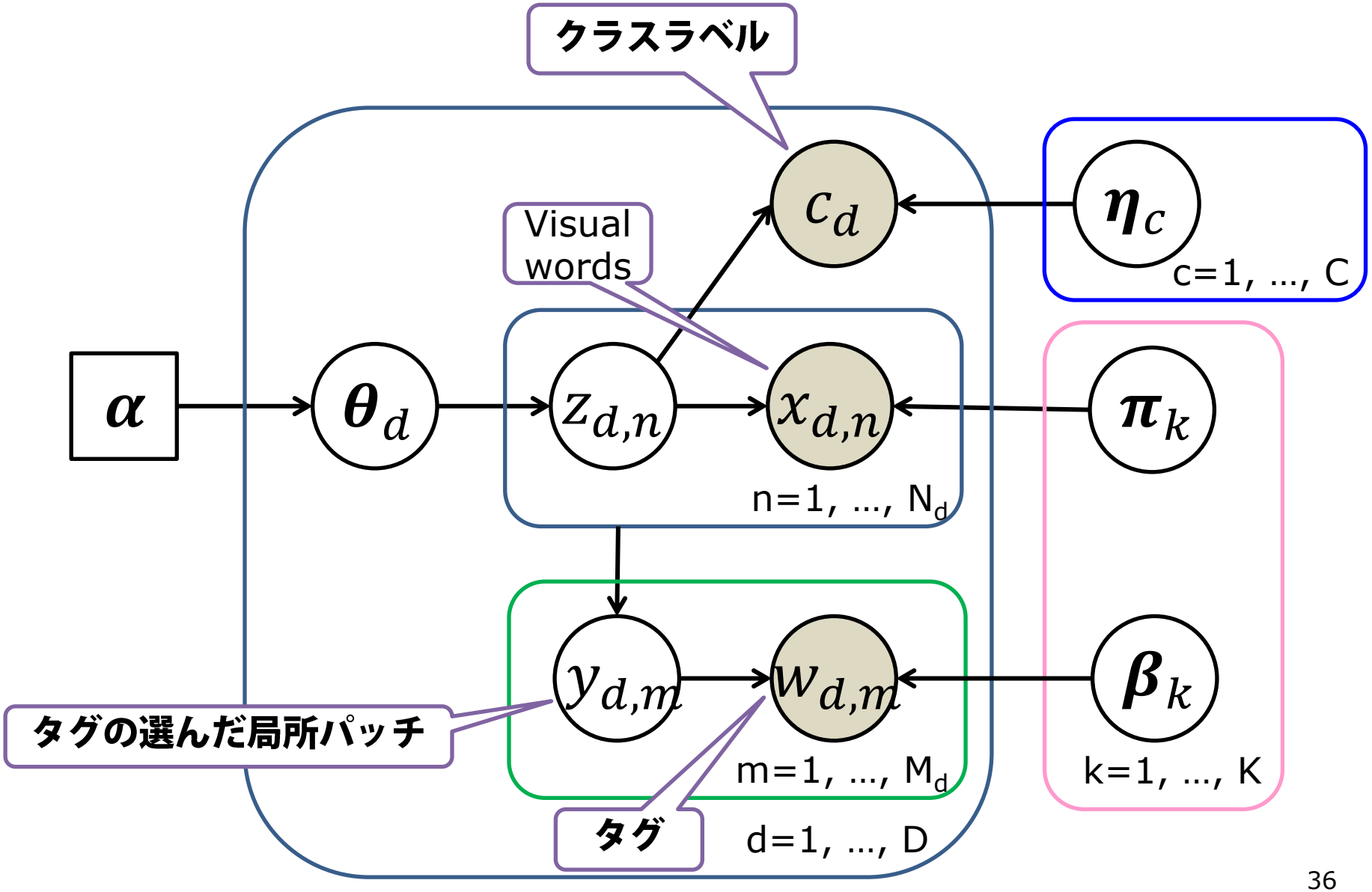
tree	.05
forest	.04
…	…

LDA



[石黒 & 竹内, 2012]

Multi-class sLDA with annotation



for 画像d = 1, 2, …, D

topic proportion $\boldsymbol{\theta}_d|\boldsymbol{\alpha}\sim\mathrm{Dir}(\boldsymbol{\alpha})$

for 単語n = 1, 2, …, N_d

topic-VW assignment $z_{d,n}|\boldsymbol{\theta}_d\sim\mathrm{Mult}(\boldsymbol{\theta}_d)$

VW observation $x_{d,n}|z_{d,n},\{\boldsymbol{\pi}_k\}\sim\mathrm{Mult}\left(\boldsymbol{\pi}_{z_{d,n}}\right)$

for タグm = 1, 2, …, M_d

tag-patch assignment $y_{d,m}\sim\mathrm{Uniform}(\{1,2,\cdots,N_d\})$

word observation $w_{d,m}|y_{d,m},\{z_{d,n}\},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}\left(\boldsymbol{\beta}_{z_{d,y_{d,m}}}\right)$

class label $c_d|\{\boldsymbol{\eta}_c\},\{z_{d,n}\}\sim\mathrm{soft-max}\left(\{\boldsymbol{\eta}_c\},\{z_{d,n}\}\right)$

for トピック k = 1, 2, …, K

topic-VW proportion $\boldsymbol{\pi}_k$

topic-tag proportion $\boldsymbol{\beta}_k$

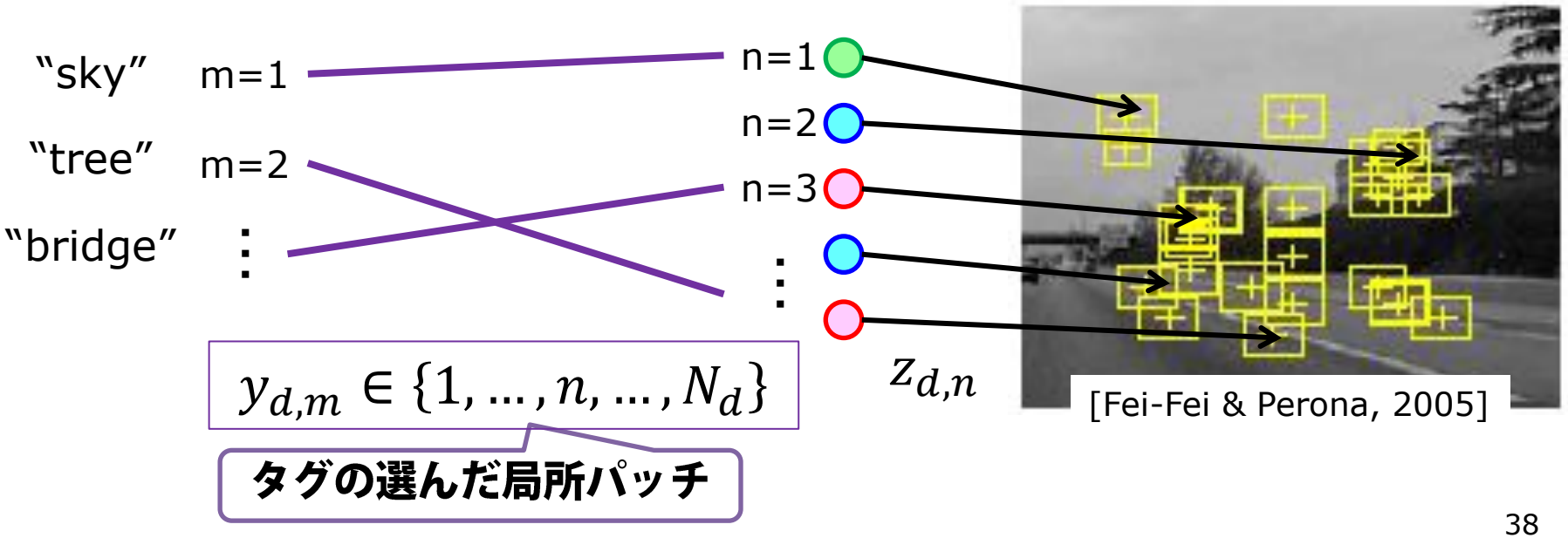
for クラスラベル c = 1, 2, …, C

class parameter for soft-max $\boldsymbol{\eta}_c$

パッチ-タグ対応

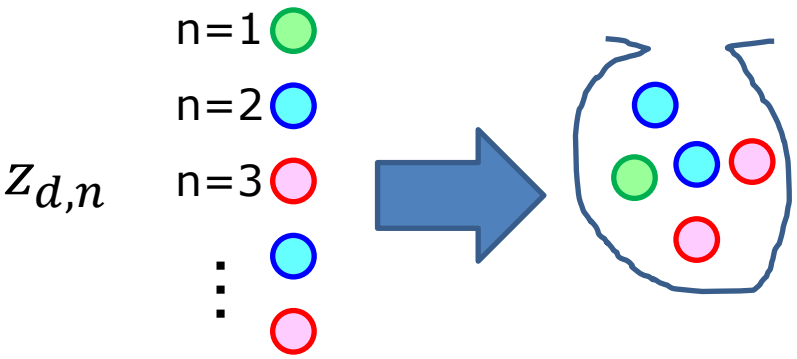
- 局所パッチとタグの対応は、何もモデルが立てられないので一様ランダムに決めます

$$y_{d,m} \sim \mathrm{Uniform}(\{1,2,\cdots,N_d\})$$



multi-class sLDAの クラスラベルモデル

- クラスラベル用のパラメータ η を使って、soft-maxによる多項分布サンプリング



文書 d の経験トピック分布

$$\bar{z}_d = \frac{1}{N_d}\sum_n z_{d,n}$$

$c_d =$ “mountain” “water” “city” “street”

$$c_d|\{\boldsymbol{\eta}_c\},\{z_{d,n}\} \sim \text{soft-max}\left(\{\boldsymbol{\eta}_c\},\{z_{d,n}\}\right)$$

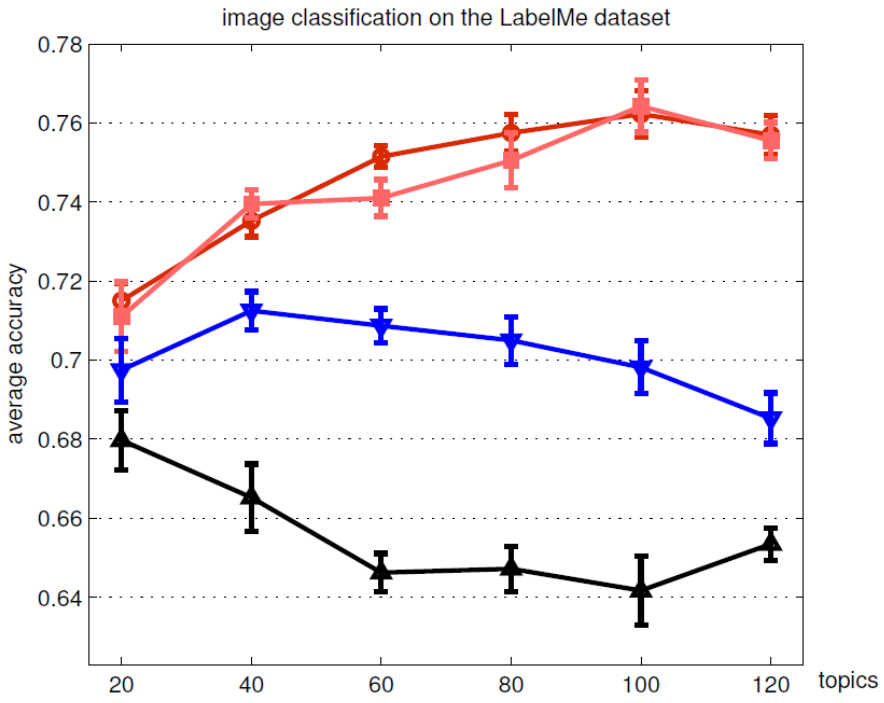
$$p\left(c_d = c|\boldsymbol{\eta}_c,\{z_{d,n}\}\right) = \frac{\exp(\boldsymbol{\eta}_c{}^T\bar{\boldsymbol{z}}_d)}{\sum_{l=1}^{C}\exp(\boldsymbol{\eta}_l{}^T\bar{\boldsymbol{z}}_d)}$$

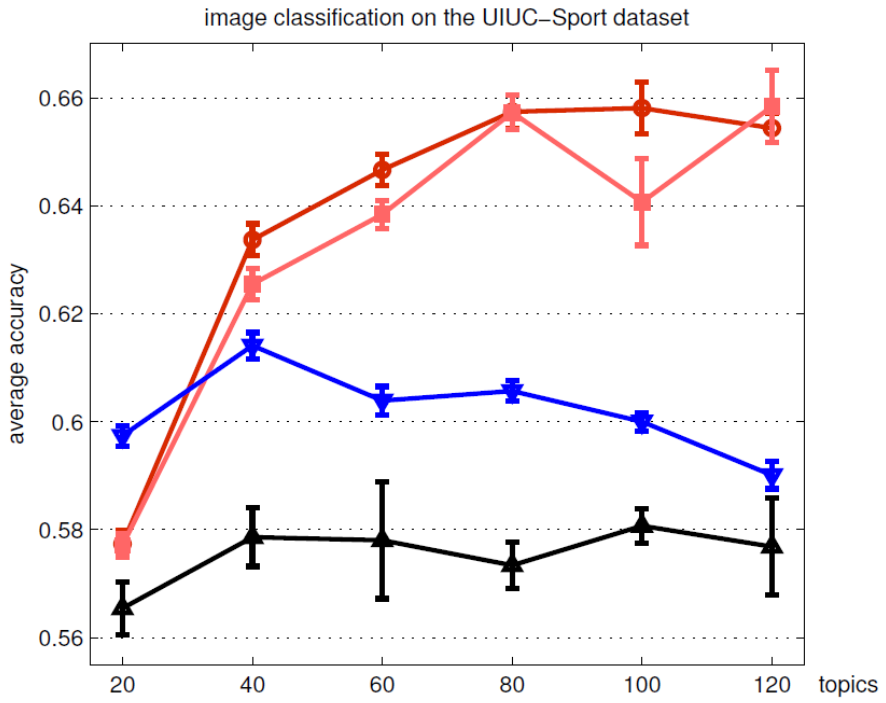
パラメータ、隠れ変数の推定

- 論文では変分ベイズ(VB)による推定法が紹介されています
- ☹ 少々複雑になります・・・

未知画像の識別

- 学習し終わったモデルに対して、未知画像 d のクラスラベル c とタグ w を推定します
- まず、画像 d の変分事後トピック分布を計算します
- これを用いて、cとwを選びますが、正確な計算は不可能なので、色々近似が必要です。
- 詳しくは論文を読んでください





[Wang, 2009]

赤: 提案法(Multi-class sLDA with annotation)

ピンク: 提案法(Multi-class sLDA)

青: Fei-Fei & Perona, 2005 (各クラスごとに個別にLDAを学習)

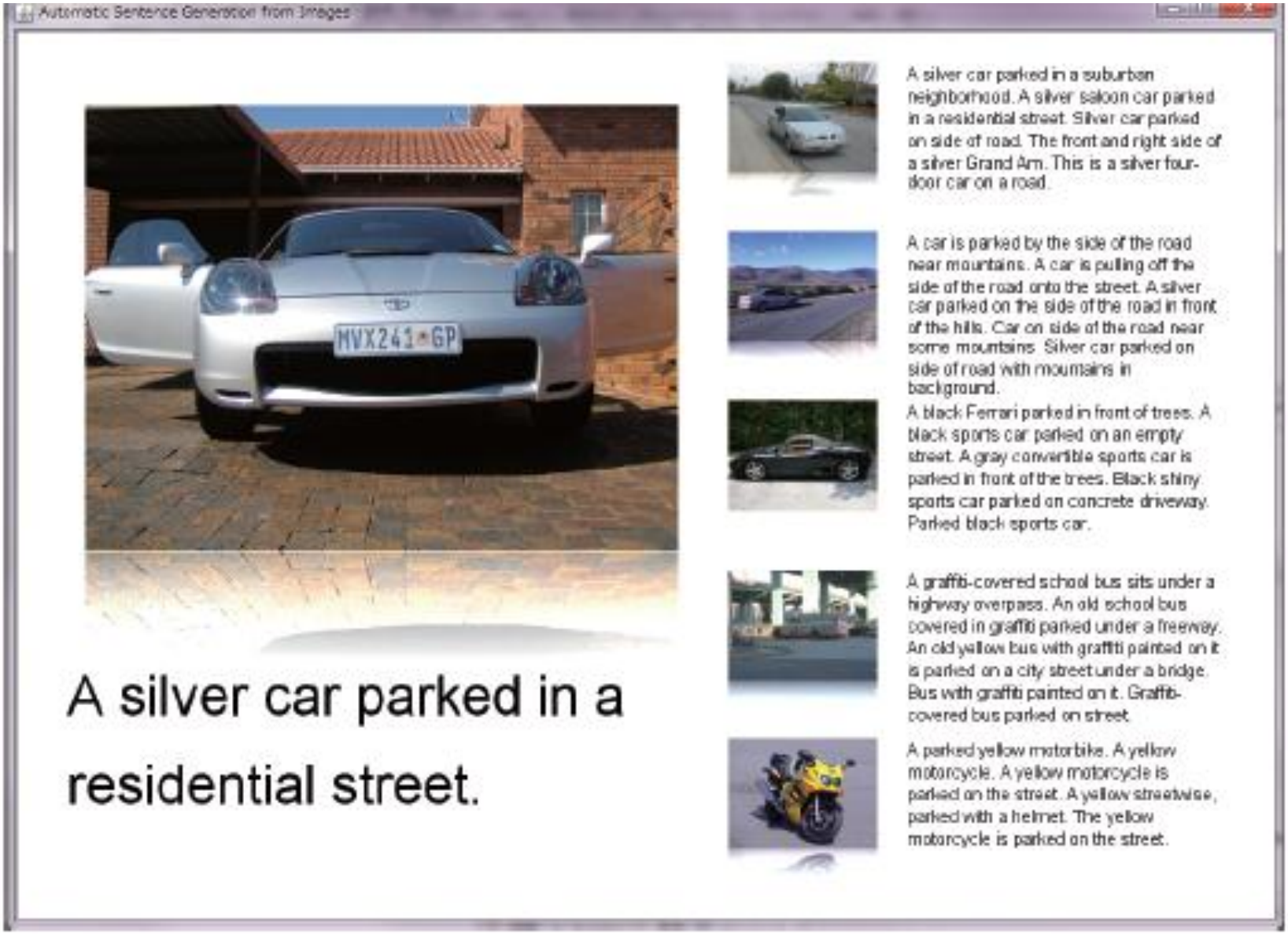
黒: Bosch, 2006 (LDA + kNN)

Correct classification with predicted annotations		**Incorrect classification (correct class) with predicted annotations**	
highway car, sign, road			*coast (highway)* car, sand beach, tree
inside city buildings, car, sidewalk			*street (inside city)* window, tree, building occluded
tall building trees, buildings occluded, window			*inside city (tall building)* tree, car, sidewalk

[Wang, 2009]

street tree, car, sidewalk			*highway (street)* car, window, tree
forest tree trunk, trees, ground grass			*mountain (forest)* snowy mountain, tree trunk
coast sand beach, cloud			*open country (coast)* sea water, buildings
mountain snowy mountain, sea water, field			*highway (mountain)* tree, snowy mountain

[Wang, 2009]


Automatic Sentence Generation from Images
A silver car parked in a residential street.
A silver car parked in a suburban neighborhood. A silver saloon car parked in a residential street. Silver car parked on side of road. The front and right side of a silver Grand Am. This is a silver four-door car on a road.
A car is parked by the side of the road near mountains. A car is pulling off the side of the road onto the street. A silver car parked on the side of the road in front of the hills. Car on side of the road near some mountains. Silver car parked on side of road with mountains in background.
A black Ferrari parked in front of trees. A black sports car parked on an empty street. A gray convertible sports car is parked in front of the trees. Black shiny sports car parked on concrete driveway. Parked black sports car.
A graffiti-covered school bus sits under a highway overpass. An old school bus covered in graffiti parked under a freeway. An old yellow bus with graffiti painted on it is parked on a city street under a bridge. Bus with graffiti painted on it. Graffiti-covered bus parked on street.
A parked yellow motorbike. A yellow motorcycle. A yellow motorcycle is parked on the street. A yellow streetwise, parked with a helmet. The yellow motorcycle is parked on the street.

[Ushiku, 2011]

まとめ: Multi-class sLDA with annotation

- BoVWベースのトピックモデルで、画像のクラスラベル、さらにタグ（アノテーション）付与まで同時にモデル化

- 画像のトピック分布から直接クラス識別
- 局所パッチからタグを生成

- アノテーション無しのMulti-class sLDA自体も多クラス識別モデルとして新規性あり
- ただのBoVW-LDAよりもはるかに高い性能

他の（動）画像応用

- BoVW: 多すぎてフォローできません

- Niebles et al., “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words”, Int. J. Computer Vision, Vol. 79, pp. 299-318, 2008.

- Rodriguez et al., “Tracking in Unstructured Crowded Scenes”, in Proc. ICCV, 2009.

- Sivic et al., “Unsupervised Discovery of Visual Object Class Hierarchies”, in Proc. CVPR, 2008.

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [Shimizu, 2008] Shimizu et al, “Super-Resolution from Image Sequence under Influence of Hot-Air Optical Turbulence”, in Proc. CVPR, 2008.
- [Viola & Jones, 2001] Viola and Jones, “Robust Real-time Object Detection”, in Proc. CVPR, 2001.
- [Lowe, 2004] Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, Vol. 60, pp. 91-110, 2004.
- [藤吉, 2007] "Gradientベースの特徴抽出 - SIFTとHOG - ", 情報処理学会 研究報告 CVIM 160, pp. 211-224, 2007.
- [Fei-Fei & Perona, 2005] Fei-Fei and Perona, “A Bayesian Hierarchical Model for Learning Natural Scene Categories”, in Proc. CVPR. 2005.

引用及び参考文献

- [Treisman & Gelade, 1980] Treisman and Gelade, “A feature-integration theory of attention”, *Cognitive Psychology*, Vol. 12. pp. 97–136, 1980.
- [Li, 2002] Li et al., “Natural scene categorization in the near absence of attention”, PNAS, Vol. 99, No. 14. pp.9596–9601, 2002.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.
- [Wang, 2009] Wang et al., “Simultaneous Image Classification and Annotation”, in Proc. CVPR, 2009.
- [Ushiku, 2011] Ushiku et al., “Understanding Images with Natural Sentences”, in Proc. ACM Multimedia, 2011.