

# **Semestrální projekt**

*Bag of Features*

Davydov Oleh a Umerenkov Valerii

# 1. Popis projektu

Cílem projektu bylo navrhnut, implementovat a prakticky demonstrovat kompletní vyhledávací pipeline založenou na modelu Bag of Features (BoF). Tento klasický přístup převádí vizuální obsah obrázku do číselného vektoru podobně, jako se v textovém vyhledávání pracuje s bag-of-words reprezentací dokumentů. Projekt tak ukazuje, jak lze principy vektorového modelu, TF-IDF váhování a kosinové podobnosti, známé z textové retrieval, aplikovat v oblasti multimediálních dat.

Systém umožňuje uživatelsky přívětivé vyhledávání podobných obrázků prostřednictvím webového rozhraní. Backend provádí samotné zpracování dotazu a porovnání s indexem, zatímco frontend poskytuje jednoduché rozhraní pro nahrání obrázku a vizualizaci výsledků.

## Vstupy projektu

- obrázek nahraný uživatelem přes webové rozhraní,
- databáze indexovaných obrázků (Oxford5k) reprezentovaných pomocí BoF vektorů.

## Výstupy projektu

- seznam nejpodobnějších obrázků z databáze,
- podobnostní skóre vypočtené pomocí kosinové míry,
- vizuální náhledy výsledků zobrazené ve webové aplikaci.

## Stručný popis principu

Každý obrázek je převeden na množinu lokálních příznaků (SIFT), které jsou následně mapovány na předem naučený vizuální slovník vytvořený metodou k-means. Výsledkem je histogram výskytu vizuálních slov vážený metodou TF-IDF. Tyto vektory mají pevnou délku a umožňují efektivní porovnávání obrázků pomocí kosinové podobnosti.

Projekt tak demonstruje klasický pipeline vizuálního vyhledávání založený na lokálních příznacích a ukazuje, že obdobné principy jako v textovém vyhledávání lze úspěšně použít i v oblasti analýzy obrazového obsahu.

## 2. Způsob řešení

Řešení projektu vychází z klasického modelu Bag of Features (BoF), který převádí obrázek na histogram výskytu tzv. vizuálních slov. Tento přístup analogicky odpovídá textovému vektorovému modelu, kde dokument reprezentujeme pomocí bag-of-words. Projekt kombinuje principy lokálních příznaků (SIFT), klastrování (k-means) a vektorového modelu s TF-IDF váhováním a kosinovou podobností.

### 2.1 Extrakce lokálních příznaků – SIFT

Každý obrázek je převeden na množinu lokálních příznaků pomocí algoritmu **SIFT**.

SIFT poskytuje stabilní body zájmu invariantní vůči:

- rotaci,
- změně měřítka,
- částečně i vůči změnám osvětlení.

Důvod použití SIFT:

- ✓ vysoká stabilita,
- ✓ vhodnost pro vyhledávání rigidních objektů,
- ✓ klasický základ BoF přístupů.

### 2.2 Tvorba vizuálního slovníku – k-means clustering

Ze všech extrahovaných SIFT deskriptorů je vytvořen soubor trénovacích příznaků, který je klasifikován pomocí algoritmu k-means do K shluků. Každý centroid odpovídá jednomu vizuálnímu slovu.

Postup:

1. Náhodně vzorkované deskriptory z datasetu.
2. MiniBatch k-means trénink (rychlejší varianta).
3. Uložení centroidů jako vizuálního slovníku.

### 2.3 Kvantizace příznaků

Každý SIFT deskriptor je přiřazen k nejbližšímu centroidu (vizuálnímu slovu).

Výsledkem je:

- pro každý obrázek: multiset vizuálních slov,
- základ pro výpočet histogramu.

## 2.4 Tvorba histogramu a váhování TF-IDF

Histogram zachycuje četnost jednotlivých vizuálních slov pro obrázek. Aby bylo možné rozlišovat běžné a vzácné prvky, je použit **TF-IDF model** známý z textového vyhledávání:

- term frequency (TF)
- inverse document frequency (IDF)
- kombinace TF-IDF:

Histogram je následně normalizován pomocí L2 normy pro použití kosinové podobnosti.

## 2.5 Vyhledávání pomocí kosinové podobnosti

Při vyhledávání podobných obrázků se používá kosinová podobnost, která hodnotí úhel mezi dvěma vektory. Tento přístup je v souladu s vektorovým modelem, kde kosinová podobnost slouží jako standardní metrika při porovnávání dokumentů.

Pokud mají dva obrázky podobně rozložená vizuální slova, jejich vektory budou mít malý úhel, a tedy vysokou míru podobnosti.

## 2.6 Použitý dataset Oxford5k

K trénování vizuálního slovníku a testování metody byl použit standardní dataset Oxford5k, který obsahuje více než pět tisíc fotografií Oxfordských památek. Tento dataset je dlouhodobě používán k hodnocení metod image retrieval a poskytuje vhodné podmínky pro ověření kvality implementovaného systému.

# 3. Implementace

Implementace projektu byla rozdělena do několika samostatných částí, které společně tvoří ucelený systém pro vyhledávání podobných obrázků. Celé řešení je postaveno na programovacím jazyce Python, doplněném o moderní

webové technologie pro tvorbu backendu a frontendového uživatelského rozhraní.

## Backend

Backend je implementován v jazyce Python za využití frameworku FastAPI. Pro extrakci lokálních příznaků se používá knihovna OpenCV, tvorba vizuálního slovníku a numerické výpočty jsou řešeny pomocí knihoven scikit-learn a NumPy.

Server běží nad ASGI technologií a poskytuje REST API pro komunikaci s frontendem.

## Frontend

Frontend je vytvořen v Reactu s využitím build systému Vite. Webová aplikace umožňuje uživateli nahrát dotazový obrázek, odeslat jej na server a následně zobrazit získané výsledky. Komunikace s backendem probíhá přes HTTP rozhraní.

## Datové soubory a modely

Projekt pracuje s několika typy datových struktur:

- SIFT deskriptory všech obrázků datasetu,
- vizuální slovník tvořený centroidy k-means,
- IDF vektor popisující význam jednotlivých vizuálních slov, TF-IDF vektory všech obrázků v databázi (index pro vyhledávání).

## 3.2 Struktura projektu

Projekt je rozdělen do několika adresářů, které odpovídají jednotlivým krokům zpracování:

- **src/** — skripty pro extrakci SIFT, tvorbu vizuálního slovníku a výpočet histogramů
- **backend/** — implementace API serveru a vyhledávací logiky
- **frontend/** — React aplikace
- **data/** — uložené modely a indexy,
- **images/** — databázové obrázky,
- **uploads/** — dočasné soubory od uživatelských dotazů.

Každý skript v části **src** odpovídá jedné etapě pipeline Bag of Features: extrakci příznaků, učení vizuálního slovníku a tvorbě BoF reprezentací.

## **3.3 Zpracování dat a vyhledávání**

Backend implementuje veškerou logiku potřebnou pro obsahové vyhledávání:

### **1) Extrakce SIFT příznaků**

Po nahrání obrázku server vytvoří množinu lokálních deskriptorů, které reprezentují výrazné oblasti obrázku. Tyto deskriptory mají jednotnou délku a tvoří základní stavební prvky pro pozdější převod do BoF reprezentace.

### **2) Tvorba vizuálního slovníku**

Během přípravné fáze je z deskriptorů všech obrázků sestaven reprezentativní vzorek. Ten je následně shlukován metodou k-means do pevného počtu skupin. Každá skupina odpovídá jednomu vizuálnímu slovu a celý soubor centroidů tvoří vizuální slovník používaný při vyhledávání.

### **3) Kvantizace příznaků a tvorba histogramu**

Každý deskriptor obrázku je přiřazen nejbližšímu vizuálnímu slovu. Tím vzniká histogram udávající četnost jednotlivých vizuálních slov. Histogram je následně normalizován a vážen pomocí TF-IDF, což odpovídá obdobné technice ve vektorovém modelu textového vyhledávání.

### **4) Vyhledávání pomocí kosinové podobnosti**

Dotazový obrázek je převeden do stejné reprezentace jako obrázky v databázi. Následně je vyhodnocena míra podobnosti mezi dotazovým vektorem a vektory indexovaných obrázků. Podobnost je určena kosinovou mírou a výsledky jsou seřazeny od nejpodobnějších po nejméně podobné.

## **3.4 Frontend – uživatelské rozhraní**

Frontend poskytuje jednoduchou webovou aplikaci, ve které uživatel:

- nahraje obrázek z počítače,
- odešle jej na server,
- obdrží vizuální náhledy nejpodobnějších obrázků a jejich skóre.

Aplikace zobrazuje výsledky v přehledné formě a umožňuje snadnou interakci bez nutnosti instalace dalších nástrojů.

## 3.5 Běhové požadavky

Pro běh backendu je vyžadována instalace Pythonu a knihoven uvedených v seznamu závislostí. Frontend vyžaduje prostředí Node.js.

Celkově je systém nenáročný na výkon a běží na běžném osobním počítači, přičemž vyšší paměťová náročnost může vzniknout při vytváření vizuálního slovníku a ukládání TF-IDF reprezentací velkých datasetů.

## 4. Příklad výstupu

V této části je uveden příklad konkrétního dotazu a odpovídajícího výstupu systému. Cílem je ukázat praktické fungování implementovaného vyhledávání podobných obrázků pomocí modelu Bag of Features.

Uživatel nahraje libovolný obrázek, který slouží jako dotazový. Backend následně provede extrakci příznaků, kvantizaci, výpočet TF-IDF reprezentace a porovnání s databází. Výsledkem je seřazený seznam nejpodobnějších obrázků.

### Ukázka dotazů:



► Top-10 similar images



all\_souls\_000000.jpg

1



all\_souls\_000085.jpg

0.7774414759213626



all\_souls\_000054.jpg

0.7752640445144706

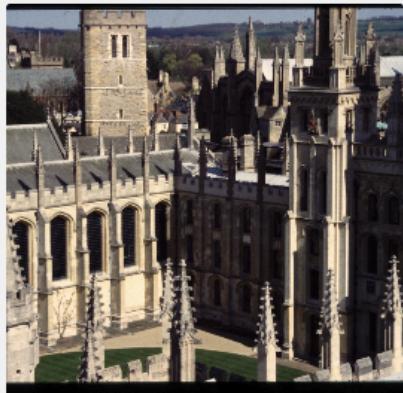


pitt\_rivers\_000138.jpg

0.7345670781381776



► Top-10 similar images



1



[0.6877486843923197](#)



[0.6750404556036185](#)



[0.6688582991959882](#)



► Top-10 similar images



worcester\_000196.jpg

[0.9874536501988097](#)



oxford\_003094.jpg

[0.8049874620344167](#)



oxford\_001204.jpg

[0.7983092920904012](#)



christ\_church\_000158.jpg

[0.7968636308589399](#)

## 5. Experimentální sekce

Cílem experimentální části bylo prověřit chování implementovaného systému při různých nastaveních klíčových parametrů pipeline Bag of Features.

Zaměřili jsme se zejména na dva faktory, které zásadně ovlivňují jak přesnost vyhledávání, tak výpočetní náročnost:

1. počet extrahovaných SIFT příznaků na obrázek,
2. velikost vizuálního slovníku (K).

Pro každý experiment byla měřena jednak přesnost systémového vyhledávání, jednak rychlosť zpracování dotazu. Hodnotili jsme typické scénáře od nízkého zatížení až po vysoké hodnoty parametrů, u nichž dochází k zahlcení systému.

**Tabulka 1 — Přesnost**

K \ nfeatures	500 SIFT	2000 SIFT	Bez omezení
300	<b>Nízká přesnost:</b> málo příznaků i malý slovník. Obrázky si nejsou dostatečně odlišeny.	<b>Lepší přesnost:</b> více příznaků zlepšuje BoW	<b>Nestabilní výsledky:</b> extrémně mnoho příznaků vede k přetížení k-means a horší kvantizaci.
700	<b>Střední přesnost:</b> slovník už dokáže rozlišit základní struktury, ale méně SIFT příznaků stále omezuje kvalitu.	<b>Dobrá přesnost —</b> nejlepší kompromis mezi velikostí slovníku a počtem příznaků; výsledky jsou konzistentní.	<b>Přesnost se nezlepšuje:</b> příliš mnoho deskriptorů vede k přeplněným histogramům a horšímu TF-IDF váhování.
1200	<b>Střední až dobrá přesnost:</b> slovník je jemný, ale málo SIFT příznaků nedokáže využít jeho kapacitu.	<b>Dobrá přesnost:</b> dobře modeluje složité scény.	<b>Může docházet ke zhoršení:</b> BoW histogramy jsou výrazně řidší a kvantizace trpí kvůli příliš velkému množství deskriptorů.

**Tabulka 2 — Rychlosť vyhledávání**

K \ nfeatures	500 SIFT	2000 SIFT	Bez omezení
300	<b>Rychlé:</b> málo příznaků, malý slovník, velmi rychlá kvantizace.	<b>Stále rychlé:</b> mírně pomalejší kvůli více příznakům.	<b>Pomalé:</b> extrakce SIFT trvá dlouho, ale kvantizace je pořád relativně rychlá.
700	<b>Rychlé:</b> stále nízká výpočetní náročnost.	<b>Středně rychlé:</b> vyvážený poměr mezi počtem centroidů a	<b>Pomalé:</b> příliš mnoho deskriptorů znatelně zvyšuje čas kvantizace.

		příznaky.	
1200	<b>Středně rychlé:</b> k-means predict zabere delší čas, ale málo příznaků to kompenzuje.	<b>Pomalejší:</b> více centroidů i příznaků prodlužuje výpočet.	<b>Nejpomalejší:</b> mnoho deskriptorů + velký slovník. Kvantizace a TF-IDF jsou výrazně náročnější.

## 6. Diskuze

Projekt Bag of Features je navržen jako demonstrace principů probíraných v přednáškách předmětu NI-VMM, zejména použití vektorového modelu, TF-IDF váhování a kosinové podobnosti v kontextu multimediálního vyhledávání. Cílem projektu nebylo vytvořit produkční systém, ale ověřit praktickou funkčnost jednotlivých kroků pipeline a porozumět jejich omezením. Tato kapitola shrnuje hlavní nedostatky zvoleného přístupu i možnosti jeho budoucího rozšíření.

### 6.1 Omezení velikosti a kvality vizuálního slovníku

Jedním z nejvýznamnějších faktorů ovlivňujících kvalitu systému je velikost vizuálního slovníku. V projektu byl použit slovník o velikosti 700 vizuálních slov, což se ukázalo jako vhodný kompromis mezi rozlišovací schopností a výpočetní náročností. Příliš malý slovník vede k tomu, že různé vizuální struktury jsou mapovány na stejná slova, což snižuje přesnost. Naopak příliš velký slovník způsobuje rozptýlení deskriptorů do mnoha clusterů a vede k méně stabilní kvantizaci.

### 6.2 Citlivost na parametry SIFT a clusteringu

Výsledná kvalita BoF reprezentace výrazně závisí na:

- počtu extrahovaných SIFT příznaků,
- kvalitě detekovaných zájmových bodů,
- způsobu vzorkování příznaků pro trénink slovníku,
- parametrech k-means (počet iterací, inicializace, velikost dávky).

Příliš vysoký počet příznaků na obrázek může vést k zahlcení clusteringu i kvantizace. Příliš nízký počet naopak nezachytí všechny relevantní části scény. Stejně tak náhodný výběr trénovacích deskriptorů může způsobit, že slovník nebude odpovídat celé variabilitě datasetu.

## 6.3 Absence geometrické verifikace

Metoda Bag of Features pracuje výhradně s histogramovou reprezentací a ignoruje prostorové uspořádání příznaků. To znamená, že:

- systém nerozlišuje mezi shodnými lokálními strukturami v různých částech obrázku,
- může dojít k falešným pozitivním shodám, pokud mají obrázky podobné lokální textury, ale zcela odlišnou globální strukturu.

## 6.4 Omezení TF-IDF reprezentace

Ačkoli TF-IDF významně zlepšuje rozlišovací schopnost histogramů, má několik omezení:

- Penalizuje velmi častá vizuální slova, což je správné u textů, ale u obrazů mohou být běžné struktury (např. hrany) stále významné.
- Předpokládá, že všechna vizuální slova jsou nezávislá, což v praxi neplatí — vizuální rysy často tvoří vzory nebo se vyskytují společně.
- Není schopna modelovat jemnou variabilitu uvnitř clusterů, která se ztrácí během kvantizace.

## 6.5 Shrnutí diskuze

Projekt úspěšně demonstруje principy klasického přístupu Bag of Features, ale zároveň ukazuje limity tohoto modelu v kontextu moderních obrazových retrieval systémů. Přes omezení zvolených metod je výsledná implementace funkční, přehledná a vhodná jako výukový základ pro hlubší pochopení problematiky.

# 7. Závěr

Cílem projektu bylo navrhnut, implementovat a experimentálně ověřit klasický přístup k vyhledávání podobných obrázků založený na modelu **Bag of Features**, který vychází z principů vektorového modelu a TF-IDF váhování používaného v textovém vyhledávání. Projekt si kládal za úkol převést tyto principy do oblasti multimediálních dat, kde jsou dokumenty nahrazeny obrázky a textová slova lokálními vizuálními příznaky.

V rámci implementace byla vytvořena kompletní pipeline zahrnující extrakci SIFT příznaků, tvorbu vizuálního slovníku metodou k-means, kvantizaci

příznaků do vizuálních slov, sestavení histogramů a jejich následné váhování pomocí TF-IDF. Tato reprezentace byla použita pro výpočet kosinové podobnosti mezi dotazovým obrázkem a obrázky v databázi. Celý systém byl integrován do webové aplikace složené z backendu (FastAPI) a frontendového uživatelského rozhraní (React).

Experimentální část ukázala, že zvolený přístup je schopen úspěšně identifikovat podobné obrázky na základě sdílených vizuálních rysů. Výsledky rovněž potvrdily teoretické předpoklady týkající se vlivu velikosti vizuálního slovníku a počtu příznaků na přesnost i rychlosť vyhledávání. Realizovaná implementace tedy splnila svůj vzdělávací účel: poskytuje funkční demonstraci klasického modelu vizuálního vyhledávání a umožňuje praktické pochopení jednotlivých kroků pipeline Bag of Features.

Projekt zároveň poukázal na limity tohoto tradičního přístupu a tím otevírá cestu k možným budoucím rozšířením, například začlenění geometrické verifikace nebo využití moderních metod založených na hlubokých neuronových sítích. Přesto lze konstatovat, že výsledná implementace naplnila zadání i cíle projektu a poskytuje ucelený základ pro další studium a experimentování v oblasti multimedialního vyhledávání.